

Yangon University of Economics

Post Graduate Diploma in Research Studies

(9th Batch)



DRS-111 Data Structure and Data Collection Methods

First Quarter

1. The Nature of Probability and Statistics

2. Frequency distributions and Graphs

6. Methods of data collection

7. Measurements in Mx and Scaling techniques

8. Questionnaire design and fieldwork plan

9. Sampling design: Theory and Practice

The Nature of Probability and Statistics

STATISTICS TODAY

Is Higher Education “Going Digital”?

Today many students take college courses online and use eBooks. Also, many students use a laptop, smartphone, or computer tablet in the classroom. With the increased use of technology, some questions about the effectiveness of this technology have been raised. For example,

How many colleges and universities offer online courses?

Do students feel that the online courses are equal in value to the traditional classroom presentations?

Approximately how many students take online courses now?

Will the number of students who take online courses increase in the future?

Has plagiarism increased since the advent of computers and the Internet?

Do laptops, smartphones, and tablets belong in the classroom?

Have colleges established any guidelines for the use of laptops, smartphones, and tablets?

To answer these questions, Pew Research Center conducted a study of college graduates and college presidents in 2011. The procedures they used and the results of the study are explained in this chapter. See Statistics Today—Revisited at the end of the chapter.



© Shutterstock/Monkey Business Images RF

OUTLINE

Introduction

- 1-1** Descriptive and Inferential Statistics
- 1-2** Variables and Types of Data
- 1-3** Data Collection and Sampling Techniques
- 1-4** Experimental Design
- 1-5** Computers and Calculators
- Summary

OBJECTIVES

After completing this chapter, you should be able to:

- 1** Demonstrate knowledge of statistical terms.
- 2** Differentiate between the two branches of statistics.
- 3** Identify types of data.
- 4** Identify the measurement level for each variable.
- 5** Identify the four basic sampling techniques.
- 6** Explain the difference between an observational and an experimental study.
- 7** Explain how statistics can be used and misused.
- 8** Explain the importance of computers and calculators in statistics.

Introduction

You may be familiar with probability and statistics through radio, television, newspapers, and magazines. For example, you may have read statements like the following found in newspapers.

- A recent survey found that 76% of the respondents said that they lied regularly to their friends.
- *The Tribune Review* reported that the average hospital stay for circulatory system ailments was 4.7 days and the average of the charges per stay was \$52,574.
- Equifax reported that the total amount of credit card debt for a recent year was \$642 billion.
- A report conducted by the SAS Holiday Shopping Styles stated that the average holiday shopper buys gifts for 13 people.
- The U.S. Department of Agriculture reported that a 5-foot 10-inch person who weighs 154 pounds will burn 330 calories for 1 hour of dancing.
- The U.S. Department of Defense reported for a recent year that the average age of active enlisted personnel was 27.4 years.

Unusual Stats

Of people in the United States, 14% said that they feel happiest in June, and 14% said that they feel happiest in December.

Interesting Fact

Every day in the United States about 120 golfers claim that they made a hole-in-one.

Historical Note

A Scottish landowner and president of the Board of Agriculture, Sir John Sinclair introduced the word *statistics* into the English language in the 1798 publication of his book on a statistical account of Scotland. The word *statistics* is derived from the Latin word *status*, which is loosely defined as a statesman.

Statistics is used in almost all fields of human endeavor. In sports, for example, a statistician may keep records of the number of yards a running back gains during a football game, or the number of hits a baseball player gets in a season. In other areas, such as public health, an administrator might be concerned with the number of residents who contract a new strain of flu virus during a certain year. In education, a researcher might want to know if new methods of teaching are better than old ones. These are only a few examples of how statistics can be used in various occupations.

Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

There are several reasons why you should study statistics.

1. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
2. You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
3. You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.

It is the purpose of this chapter to introduce the goals for studying statistics by answering questions such as the following:

What are the branches of statistics?

What are data?

How are samples selected?

1-1 Descriptive and Inferential Statistics

OBJECTIVE 1

Demonstrate knowledge of statistical terms.

Historical Note

The 1880 Census had so many questions on it that it took 10 years to publish the results.

Historical Note

The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 B.C. In addition, the Roman Emperor Augustus (27 B.C.–A.D. 17) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly.

OBJECTIVE 2

Differentiate between the two branches of statistics.

To gain knowledge about seemingly haphazard situations, statisticians collect information for *variables*, which describe the situation.

A **variable** is a characteristic or attribute that can assume different values.

Data are the values (measurements or observations) that the variables can assume. Variables whose values are determined by chance are called **random variables**.

Suppose that an insurance company studies its records over the past several years and determines that, on average, 3 out of every 100 automobiles the company insured were involved in accidents during a 1-year period. Although there is no way to predict the specific automobiles that will be involved in an accident (random occurrence), the company can adjust its rates accordingly, since the company knows the general pattern over the long run. (That is, on average, 3% of the insured automobiles will be involved in an accident each year.)

A collection of data values forms a **data set**. Each value in the data set is called a **data value** or a **datum**.

In statistics it is important to distinguish between a sample and a population.

A **population** consists of all subjects (human or otherwise) that are being studied.

When data are collected from every subject in the population, it is called a *census*.

For example, every 10 years the United States conducts a census. The primary purpose of this census is to determine the apportionment of the seats in the House of Representatives.

The first census was conducted in 1790 and was mandated by Article 1, Section 2 of the Constitution. As the United States grew, the scope of the census also grew. Today the Census limits questions to populations, housing, manufacturing, agriculture, and mortality. The Census is conducted by the Bureau of the Census, which is part of the Department of Commerce.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

A **sample** is a group of subjects selected from a population.

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population. See Figure 1-1.

However, the information obtained from a statistical sample is said to be *biased* if the results from the sample of a population are radically different from the results of a census of the population. Also, a sample is said to be biased if it does not represent the population from which it has been selected. The techniques used to properly select a sample are explained in Section 1-3.

The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are

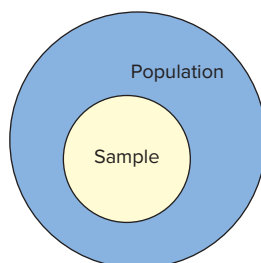
1. Descriptive statistics
2. Inferential statistics

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

In *descriptive statistics* the statistician tries to describe a situation. Consider the national census conducted by the U.S. government every 10 years. Results of this census give you the average age, income, and other characteristics of the U.S. population. To obtain this information, the Census Bureau must have some means to collect relevant data. Once data are collected, the bureau must organize and summarize them. Finally, the bureau needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

FIGURE 1-1

Population and Sample



The second area of statistics is called *inferential statistics*.

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Historical Note

Inferential statistics originated in the 1600s, when John Graunt published his book on population growth, *Natural and Political Observations Made upon the Bills of Mortality*. About the same time, another mathematician/astronomer, Edmond Halley, published the first complete mortality tables. (Insurance companies use mortality tables to determine life insurance rates.)

Unusual Stat

Twenty-nine percent of Americans want their boss's job.

Here, the statistician tries to make inferences from *samples* to *populations*. Inferential statistics uses **probability**, i.e., the chance of an event occurring. You may be familiar with the concepts of probability through various forms of gambling. If you play cards, dice, bingo, or lotteries, you win or lose according to the laws of probability. Probability theory is also used in the insurance industry and other areas.

The area of inferential statistics called **hypothesis testing** is a decision-making process for evaluating claims about a population, based on information obtained from samples. For example, a researcher may wish to know if a new drug will reduce the number of heart attacks in men over age 70 years of age. For this study, two groups of men over age 70 would be selected. One group would be given the drug, and the other would be given a placebo (a substance with no medical benefits or harm). Later, the number of heart attacks occurring in each group of men would be counted, a statistical test would be run, and a decision would be made about the effectiveness of the drug.

Statisticians also use statistics to determine *relationships* among variables. For example, relationships were the focus of the most noted study in the 20th century, "Smoking and Health," published by the Surgeon General of the United States in 1964. He stated that after reviewing and evaluating the data, his group found a definite relationship between smoking and lung cancer. He did not say that cigarette smoking actually causes lung cancer, but that there is a relationship between smoking and lung cancer. This conclusion was based on a study done in 1958 by Hammond and Horn. In this study, 187,783 men were observed over a period of 45 months. The death rate from lung cancer in this group of volunteers was 10 times as great for smokers as for nonsmokers.

Finally, by studying past and present data and conditions, statisticians try to make predictions based on this information. For example, a car dealer may look at past sales records for a specific month to decide what types of automobiles and how many of each type to order for that month next year.

EXAMPLE 1-1 Descriptive or Inferential Statistics

Determine whether descriptive or inferential statistics were used.

- The average price of a 30-second ad for the Academy Awards show in a recent year was 1.90 million dollars.
- The Department of Economic and Social Affairs predicts that the population of Mexico City, Mexico, in 2030 will be 238,647,000 people.
- A medical report stated that taking statins is proven to lower heart attacks, but some people are at a slightly higher risk of developing diabetes when taking statins.
- A survey of 2234 people conducted by the Harris Poll found that 55% of the respondents said that excessive complaining by adults was the most annoying social media habit.

SOLUTION

- A descriptive statistic (average) was used since this statement was based on data obtained in a recent year.
- Inferential statistics were used since this is a prediction for a future year.
- Inferential statistics were used since this conclusion was drawn from data obtained from samples and used to conclude that the results apply to a population.
- Descriptive statistics were used since this is a result obtained from a sample of 2234 survey respondents.

Applying the Concepts 1–1

Attendance and Grades

Read the following on attendance and grades, and answer the questions.

A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.

Based on this information, attendance and grades are related. The more you attend class, the more likely it is you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

1. What are the variables under study?
2. What are the data in the study?
3. Are descriptive, inferential, or both types of statistics used?
4. What is the population under study?
5. Was a sample collected? If so, from where?
6. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Unusual Stat

Only one-third of crimes committed are reported to the police.

Exercises 1–1

1. Define statistics.
2. What is a variable?
3. What is meant by a census?
4. How does a population differ from a sample?
5. Explain the difference between descriptive and inferential statistics.
6. Name three areas where probability is used.
7. Why is information obtained from samples used more often than information obtained from populations?
8. What is meant by a biased sample?
9. Because of the current economy, 49% of 18- to 34- year-olds have taken a job to pay the bills. (Source: Pew Research Center)
10. In 2025, the world population is predicted to be 8 billion people. (Source: United Nations)
11. In a weight loss study using teenagers at Boston University, 52% of the group said that they lost weight and kept it off by counting calories.
12. Based on a sample of 2739 respondents, it is estimated that pet owners spent a total of 14 billion dollars on veterinarian care for their pets. (Source: American Pet Products Association, Pet Owners Survey)
13. A recent article stated that over 38 million U.S. adults binge-drink alcohol.
14. The Centers for Disease Control and Prevention estimated that for a specific school year, 7% of children in kindergartens in the state of Oregon had nonmedical waivers for vaccinations.
15. A study conducted by a research network found that people with fewer than 12 years of education had lower life expectancies than those with more years of education.
16. A survey of 1507 smartphone users showed that 38% of them purchased insurance at the same time as they purchased their phones.
17. Forty-four percent of the people in the United States have type O blood. (Source: American Red Cross)

For Exercises 9–17, determine whether descriptive or inferential statistics were used.

Extending the Concepts

18. Find three statistical studies and explain whether they used descriptive or inferential statistics.

19. Find a gambling game and explain how probability was used to determine the outcome.

1-2 Variables and Types of Data

OBJECTIVE 3

Identify types of data.

As stated in Section 1-1, statisticians gain information about a particular situation by collecting data for random variables. This section will explore in greater detail the nature of variables and types of data.

Variables can be classified as qualitative or quantitative.

Qualitative variables are variables that have distinct categories according to some characteristic or attribute.

For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative. Other examples of qualitative variables are religious preference and geographic locations.

Quantitative variables are variables that can be counted or measured.

For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

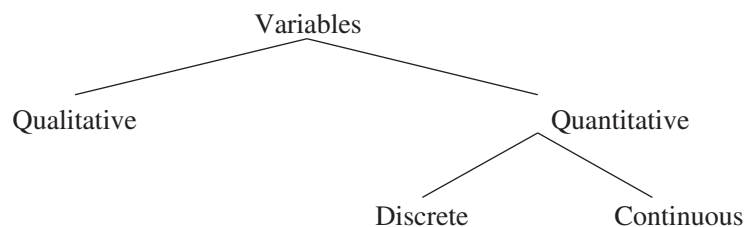
Quantitative variables can be further classified into two groups: discrete and continuous. *Discrete variables* can be assigned values such as 0, 1, 2, 3 and are said to be *countable*. Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a call center each day for a month.

Discrete variables assume values that can be counted.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

The classification of variables can be summarized as follows:



EXAMPLE 1–2 Discrete or Continuous Data

Classify each variable as a discrete or continuous variable.

- The number of hours during a week that children ages 12 to 15 reported that they watched television.
- The number of touchdowns a quarterback scored each year in his college football career.
- The amount of money a person earns per week working at a fast-food restaurant.
- The weights of the football players on the teams that play in the NFL this year.

SOLUTION

- Continuous, since the variable time is measured
- Discrete, since the number of touchdowns is counted
- Discrete, since the smallest value that money can assume is in cents
- Continuous, since the variable weight is measured

Unusual Stat

Fifty-two percent of Americans live within 50 miles of a coastal shoreline.

Since continuous data must be measured, answers must be rounded because of the limits of the measuring device. Usually, answers are rounded to the nearest given unit. For example, heights might be rounded to the nearest inch, weights to the nearest ounce, etc. Hence, a recorded height of 73 inches could mean any measure from 72.5 inches up to but not including 73.5 inches. Thus, the boundary of this measure is given as 72.5–73.5 inches. The **boundary** of a number, then, is defined as a class in which a data value would be placed before the data value was rounded. *Boundaries are written for convenience as 72.5–73.5 but are understood to mean all values up to but not including 73.5.* Actual data values of 73.5 would be rounded to 74 and would be included in a class with boundaries of 73.5 up to but not including 74.5, written as 73.5–74.5. As another example, if a recorded weight is 86 pounds, the exact boundaries are 85.5 up to but not including 86.5, written as 85.5–86.5 pounds. Table 1–1 helps to clarify this concept. The boundaries of a continuous variable are given in one additional decimal place and always end with the digit 5.

TABLE 1–1 Recorded Values and Boundaries

Variable	Recorded value	Boundaries
Length	15 centimeters (cm)	14.5–15.5 cm
Temperature	86 degrees Fahrenheit (°F)	85.5–86.5°F
Time	0.43 second (sec)	0.425–0.435 sec
Mass	1.6 grams (g)	1.55–1.65 g

EXAMPLE 1–3 Class Boundaries

Find the boundaries for each measurement.

- 17.6 inches
- 23° Fahrenheit
- 154.62 mg/dl

SOLUTION

- 17.55–18.55 inches
- 22.5–23.5° Fahrenheit
- 154.615–154.625 mg/dl

OBJECTIVE 4

Identify the measurement level for each variable.

Unusual Stat

Sixty-three percent of us say we would rather hear the bad news first.

Historical Note

When data were first analyzed statistically by Karl Pearson and Francis Galton, almost all were continuous data. In 1899, Pearson began to analyze discrete data. Pearson found that some data, such as eye color, could not be measured, so he termed such data *nominal data*. Ordinal data were introduced by a German numerologist Frederick Mohs in 1822 when he introduced a hardness scale for minerals. For example, the hardest stone is the diamond, which he assigned a hardness value of 1500. Quartz was assigned a hardness value of 100. This does not mean that a diamond is 15 times harder than quartz. It only means that a diamond is harder than quartz. In 1947, a psychologist named Stanley Smith Stevens made a further division of continuous data into two categories, namely, interval and ratio.

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.? Or are the values obtained from measurement, such as heights, IQs, or temperature? This type of classification—i.e., how variables are categorized, counted, or measured—uses **measurement scales**, and four common types of scales are used: nominal, ordinal, interval, and ratio.

The first level of measurement is called the *nominal level* of measurement. A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics) is an example of nominal-level measurement. Classifying survey subjects as male or female is another example of nominal-level measurement. No ranking or order can be placed on the data. Classifying residents according to zip codes is also an example of the nominal level of measurement. Even though numbers are assigned as zip codes, there is no meaningful order or ranking. Other examples of nominal-level data are political party (Democratic, Republican, independent, etc.), religion (Christianity, Judaism, Islam, etc.), and marital status (single, married, divorced, widowed, separated).

The **nominal level of measurement** classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data.

The next level of measurement is called the *ordinal level*. Data measured at this level can be placed into categories, and these categories can be ordered, or ranked. For example, from student evaluations, guest speakers might be ranked as superior, average, or poor. Floats in a homecoming parade might be ranked as first place, second place, etc. *Note that precise measurement of differences in the ordinal level of measurement does not exist.* For instance, when people are classified according to their build (small, medium, or large), a large variation exists among the individuals in each class.

Other examples of ordinal data are letter grades (A, B, C, D, F).

The **ordinal level of measurement** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.

The third level of measurement is called the *interval level*. This level differs from the ordinal level in that precise differences do exist between units. For example, many standardized psychological tests yield values measured on an interval scale. IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1°F between each unit, such as 72 and 73°F. *One property is lacking in the interval scale: There is no true zero.* For example, IQ tests do not measure people who have no intelligence. For temperature, 0°F does not mean no heat at all.

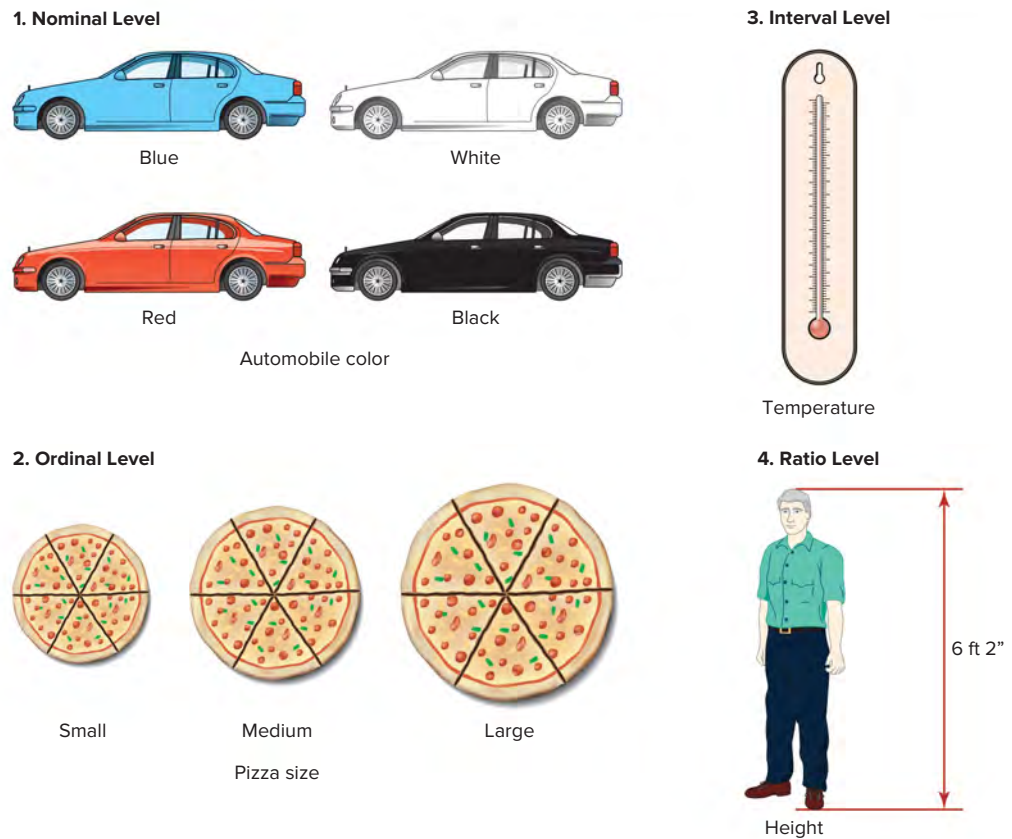
The **interval level of measurement** ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

The final level of measurement is called the *ratio level*. Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.

The **ratio level of measurement** possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

TABLE 1–2 Examples of Measurement Scales

Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level data
Zip code Gender (male, female) Eye color (blue, brown, green, hazel) Political affiliation Religious affiliation Major field (mathematics, computers, etc.) Nationality	Grade (A, B, C, D, F) Judging (first place, second place, etc.) Rating scale (poor, good, excellent) Ranking of tennis players	SAT score IQ Temperature	Height Weight Time Salary Age

FIGURE 1–2
Measurement Scales

There is not complete agreement among statisticians about the classification of data into one of the four categories. For example, some researchers classify IQ data as ratio data rather than interval. Also, data can be altered so that they fit into a different category. For instance, if the incomes of all professors of a college are classified into the three categories of low, average, and high, then a ratio variable becomes an ordinal variable. Table 1–2 gives some examples of each type of data. See Figure 1–2.

EXAMPLE 1–4 Measurement Levels

What level of measurement would be used to measure each variable?

- The ages of authors who wrote the hardback versions of the top 25 fiction books sold during a specific week
- The colors of baseball hats sold in a store for a specific year
- The highest temperature for each day of a specific month
- The ratings of bands that played in the homecoming parade at a college

SOLUTION

- a. Ratio
- b. Nominal
- c. Interval
- d. Ordinal

Applying the Concepts 1–2

Fatal Transportation Injuries

Read the following information about the number of fatal accidents for the transportation industry in for a specific year, and answer each question.

Industry	Number of fatalities
Highway accidents	968
Railway accidents	44
Water vehicle accidents	52
Aircraft accidents	151

Source: Bureau of Labor Statistics.

1. Name the variables under study.
2. Categorize each variable as quantitative or qualitative.
3. Categorize each quantitative variable as discrete or continuous.
4. Identify the level of measurement for each variable.
5. The railroad had the fewest fatalities for the specific year. Does that mean railroads have fewer accidents than the other industries?
6. What factors other than safety influence a person's choice of transportation?
7. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Exercises 1–2

1. Explain the difference between qualitative variables and quantitative variables.
2. Explain the difference between discrete and continuous variables.
3. Why are continuous variables rounded when they are used in statistical studies?
4. Name and define the four types of measurement levels used in statistics.

For Exercises 5–10, determine whether the data are qualitative or quantitative.

5. Sizes of soft drinks sold by a fast-food restaurant (small, medium, and large)
 6. Pizza sizes (small, medium, and large)
 7. Cholesterol counts for individuals
 8. Microwave wattage
 9. Number of degrees awarded by a college each year for the last 10 years
 10. Ratings of teachers
- For Exercises 11–16, determine whether the data are discrete or continuous.
11. Number of phone calls received by a 911 call center each day
 12. Systolic blood pressure readings
 13. Weights of the suitcases of airline passengers on a specific flight
 14. Votes received by mayoral candidates in a city election
 15. Number of students in the mathematics classes during the fall semester at your school for a particular school year
 16. Temperatures at a seashore resort

For Exercises 17–22, give the boundaries of each value.

17. 24 feet

18. 6.3 millimeters

19. 143 miles

20. 19.63 tons

21. 200.7 miles

22. 19 quarts

For Exercises 23–30, classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

23. Telephone numbers

24. Leap years: . . . 2016, 2020, 2024, . . .

25. Distances communication satellites in orbit are from Earth

26. Scores on a statistical final exam

27. Rating of cooked ribs at a rib cook-off

28. Blood types—O, A, B, AB

29. Online spending in dollars

30. Horsepower of automobile engines

1–3 Data Collection and Sampling Techniques

OBJECTIVE 5

Identify the four basic sampling techniques.

In research, statisticians use data in many different ways. As stated previously, data can be used to describe situations or events. For example, a manufacturer might want to know something about the consumers who will be purchasing his product so he can plan an effective marketing strategy. In another situation, the management of a company might survey its employees to assess their needs in order to negotiate a new contract with the employees' union. Data can be used to determine whether the educational goals of a school district are being met. Finally, trends in various areas, such as the stock market, can be analyzed, enabling prospective buyers to make more intelligent decisions concerning what stocks to purchase. These examples illustrate a few situations where collecting data will help people make better decisions on courses of action.

Data can be collected in a variety of ways. One of the most common methods is through the use of surveys. Surveys can be done by using a variety of methods. Three of the most common methods are the telephone survey, the mailed questionnaire, and the personal interview.

Telephone surveys have an advantage over personal interview surveys in that they are less costly. Also, people may be more candid in their opinions since there is no face-to-face contact. A major drawback to the telephone survey is that some people in the population will not have phones or will not answer when the calls are made; hence, not all people have a chance of being surveyed. Also, many people now have unlisted numbers and cell phones, so they cannot be surveyed. Finally, even the tone of voice of the interviewer might influence the response of the person who is being interviewed.

Mailed questionnaire surveys can be used to cover a wider geographic area than telephone surveys or personal interviews since mailed questionnaire surveys are less expensive to conduct. Also, respondents can remain anonymous if they desire. Disadvantages of mailed questionnaire surveys include a low number of responses and inappropriate answers to questions. Another drawback is that some people may have difficulty reading or understanding the questions.



© Banana Stock Ltd RF

Historical Note

A pioneer in census taking was Pierre-Simon de Laplace. In 1780, he developed the Laplace method of estimating the population of a country. The principle behind his method was to take a census of a few selected communities and to determine the ratio of the population to the number of births in these communities. (Good birth records were kept.) This ratio would be used to multiply the number of births in the entire country to estimate the number of citizens in the country.

Personal interview surveys have the advantage of obtaining in-depth responses to questions from the person being interviewed. One disadvantage is that interviewers must be trained in asking questions and recording responses, which makes the personal interview survey more costly than the other two survey methods. Another disadvantage is that the interviewer may be biased in his or her selection of respondents.

Data can also be collected in other ways, such as *surveying records* or *direct observation* of situations.

As stated in Section 1–1, researchers use samples to collect data and information about a particular variable from a large population. Using samples saves time and money and in some cases enables the researcher to get more detailed information about a particular subject. Remember, samples cannot be selected in haphazard ways because the information obtained might be biased. For example, interviewing people on a street corner during the day would not include responses from people working in offices at that time or from people attending school; hence, not all subjects in a particular population would have a chance of being selected.

To obtain samples that are unbiased—i.e., that give each subject in the population an equally likely chance of being selected—statisticians use four basic methods of sampling: random, systematic, stratified, and cluster sampling.

Historical Note

The first census in the United States was conducted in 1790. Its purpose was to ensure proper Congressional representation.

Random Sampling

A **random sample** is a sample in which all members of the population have an equal chance of being selected.

Random samples are selected by using chance methods or random numbers. One such method is to number each subject in the population. Then place numbered cards in a bowl, mix them thoroughly, and select as many cards as needed. The subjects whose numbers are selected constitute the sample. Since it is difficult to mix the cards thoroughly, there is a chance of obtaining a biased sample. For this reason, statisticians use another method of obtaining numbers. They generate random numbers with a computer or calculator. Before the invention of computers, random numbers were obtained from tables.

Some five-digit random numbers are shown in Table D in Appendix A. A section of Table D is shown on page 13. To select a random sample of, say, 15 subjects out of 85 subjects, it is necessary to number each subject from 01 to 85. Then select a starting number by closing your eyes and placing your finger on a number in the table. (Although this may sound somewhat unusual, it enables us to find a starting number at random.) In this case, suppose your finger landed on the number 88948 in the fourth column, the fifth number down from the top. Since you only need two-digit numbers, you can use the last two digits of each of these numbers. The first random number then is 48. Then proceed down until you have selected 15 different numbers between and including 01 and 85. When you reach the bottom of the column, go to the top of the next column. If you select a number 00 or a number greater than 85 or a duplicate number, just omit it.

In our example, we use the numbers (which correspond to the subjects) 48, 43, 44, 19, 07, 27, 58, 24, 68, and so on. Use Table D in the Appendix to get all the random numbers.

Systematic Sampling

A **systematic sample** is a sample obtained by selecting every k^{th} member of the population where k is a counting number.

Researchers obtain systematic samples by numbering each subject of the population and then selecting every k th subject. For example, suppose there were 2000 subjects in the population and a sample of 50 subjects was needed. Since $2000 \div 50 = 40$, then $k = 40$, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at random. Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until

Many overweight people have difficulty losing weight. *Prevention* magazine reported that researchers from Washington University School of Medicine studied the diets of 48 adult weight loss participants. They used food diaries, exercise monitors, and weigh-ins. They found that the participants ate an average of 236 more calories on Saturdays than they did on the other weekdays. This would amount to a weight gain of 9 pounds per year. So if you are watching your diet, be careful on Saturdays.

Are the statistics reported in this study descriptive or inferential in nature? What type of variables are used here?



© Jacobs Stock Photography/Getty Images RF

TABLE D Random Numbers

51455	02154	06955	88858	02158	76904	28864	95504	68047	41196	88582	99062	21984	67932
06512	07836	88456	36313	30879	51323	76451	25578	15986	50845	57015	53684	57054	93261
71308	35028	28065	74995	03251	27050	31692	12910	14886	85820	42664	68830	57939	34421
60035	97320	62543	61404	94367	07080	66112	56180	15813	15978	63578	13365	60115	99411
64072	76075	91393	88948	99244	60809	10784	36380	5721	24481	86978	74102	49979	28572
14914	85608	96871	74743	73692	53664	67727	21440	13326	98590	93405	63839	65974	05294
93723	60571	17559	96844	88678	89256	75120	62384	77414	24023	82121	01796	03907	35061
86656	43736	62752	53819	81674	43490	07850	61439	52300	55063	50728	54652	63307	83597
31286	27544	44129	51107	53727	65479	09688	57355	20426	44527	36896	09654	63066	92393
95519	78485	20269	64027	53229	59060	99269	12140	97864	31064	73933	37369	94656	57645
78019	75498	79017	22157	22893	88109	57998	02582	34259	11405	97788	37718	64071	66345
45487	22433	62809	98924	96769	24955	60283	16837	02070	22051	91191	40000	36480	07822
64769	25684	33490	25168	34405	58272	90124	92954	43663	39556	40269	69189	68272	60753
00464	62924	83514	97860	98982	84484	18856	35260	22370	22751	89716	33377	97720	78982
73714	36622	04866	00885	34845	26118	47003	28924	98813	45981	82469	84867	50443	00641
84032	71228	72682	40618	69303	58466	03438	67873	87487	33285	19463	02872	36786	28418
70609	51795	47988	49658	29651	93852	27921	16258	28666	41922	33353	38131	64115	39541
37209	94421	49043	11876	43528	93624	55263	29863	67709	39952	50512	93074	66938	09515
80632	65999	34771	06797	02318	74725	10841	96571	12052	41478	50020	59066	30860	96357

50 subjects were obtained. When using systematic sampling, you must be careful about how the subjects in the population are numbered. If subjects were arranged in a manner such as wife, husband, wife, husband, and every 40th subject were selected, the sample would consist of all husbands. Numbering is not always necessary. For example, a researcher may select every 10th item from an assembly line to test for defects.

Systematic sampling has the advantage of selecting subjects throughout an ordered population. This sampling method is fast and convenient if the population can be easily numbered.

Stratified Sampling

A **stratified sample** is a sample obtained by dividing the population into subgroups or strata according to some characteristic relevant to the study. (There can be several subgroups.) Then subjects are selected at random from each subgroup.

Samples within the strata should be randomly selected. For example, suppose the president of a two-year college wants to learn how students feel about a certain issue. Furthermore, the president wishes to see if the opinions of first-year students differ from those of second-year students. The president will randomly select students from each subgroup to use in the sample.

Cluster Sampling

Historical Note

In 1936, the *Literary Digest*, on the basis of a biased sample of its subscribers, predicted that Alf Landon would defeat Franklin D. Roosevelt in the upcoming presidential election. Roosevelt won by a landslide. The magazine ceased publication the following year.

A **cluster sample** is obtained by dividing the population into sections or clusters and then selecting one or more clusters at random and using all members in the cluster(s) as the members of the sample.

Here the population is divided into groups or clusters by some means such as geographic area or schools in a large school district. Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples. Suppose a researcher wishes to survey apartment dwellers in a large city. If there are 10 apartment buildings in the city, the researcher can select at random 2 buildings from the 10 and interview all the residents of these buildings. Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area. For example, if one wanted to do a study involving the patients in the hospitals in New York City, it would be very costly and time-consuming to try to obtain a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected at random, and the patients in these hospitals would be interviewed in a cluster. See Figure 1–3.

The main difference between stratified sampling and cluster sampling is that although in both types of sampling the population is divided into groups, the subjects in the groups for stratified sampling are more or less homogeneous, that is, they have similar characteristics, while the subjects in the clusters form “miniature populations.” That is, they vary in characteristics as does the larger population. For example, if a researcher wanted to use the freshman class at a university as the population, he or she might use a class of students in a freshman orientation class as a cluster sample. If the researcher were using a stratified sample, she or he would need to divide the students of the freshman class into groups according to their major field, gender, age, etc., or other samples from each group.

Cluster samples save the researcher time and money, but the researcher must be aware that sometimes a cluster does not represent the population.

The four basic sampling methods are summarized in Table 1–3.

Other Sampling Methods

Interesting Facts

Older Americans are less likely to sacrifice happiness for a higher-paying job. According to one survey, 38% of those aged 18–29 said they would choose more money over happiness, while only 3% of those over age 65 would.

In addition to the four basic sampling methods, researchers use other methods to obtain samples. One such method is called a **convenience sample**. Here a researcher uses subjects who are convenient. For example, the researcher may interview subjects entering a local mall to determine the nature of their visit or perhaps what stores they will be patronizing. This sample is probably not representative of the general customers for several reasons. For one thing, it was probably taken at a specific time of day, so not all customers entering the mall have an equal chance of being selected since they were not there when the survey was being conducted. But convenience samples can be representative of the population. If the researcher investigates the characteristics of the population and determines that the sample is representative, then it can be used.

Another type of sample that is used in statistics is a *volunteer sample* or *self-selected sample*. Here respondents decide for themselves if they wish to be included in the sample. For example, a radio station in Pittsburgh asks a question about a situation and then asks people to call one number if they agree with the action taken or call another number if they disagree with the action. The results are then announced at the end of the day. Note that most often, only people with strong opinions will call. The station does explain that this is not a “scientific poll.”

FIGURE 1-3 Sampling Methods

1. Random

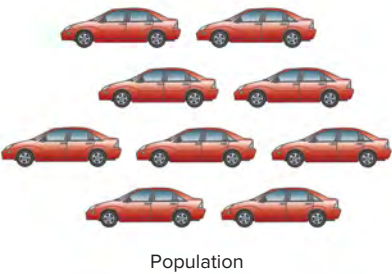
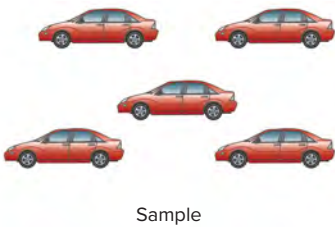
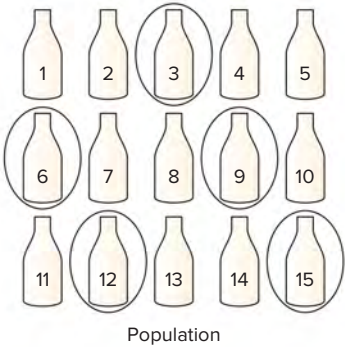


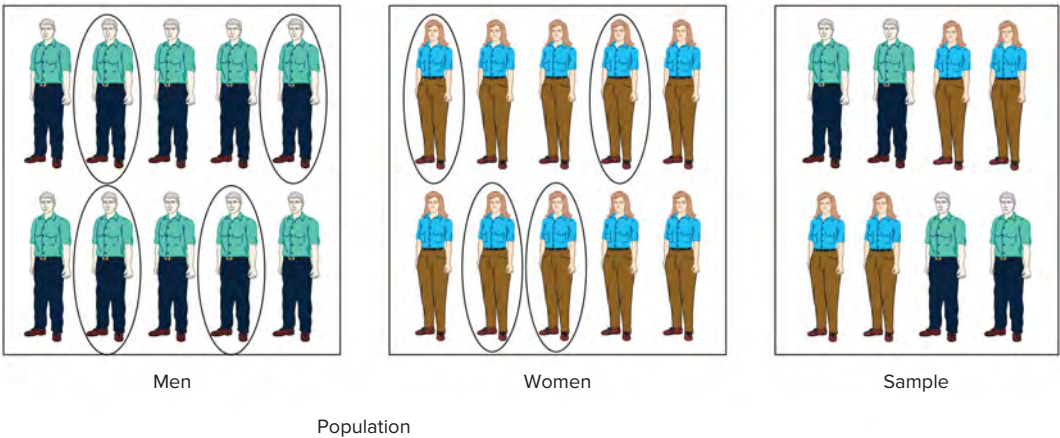
Table D		Random Numbers			
10480	15011	01536	02011	81647	
22368	46573	25595	85393	30995	
24130	48360	22527	97265	76393	
42167	93093	06243	61680	07856	
37570	39975	81837	16656	06121	
77921	06907	11008	42751	27750	
99562	72905	56420	69994	98872	
96301	91977	05463	07972	18876	
89579	14342	63661	10281	17453	
85475	36857	43342	53988		
28918	69578	88321			
63553	40961				



2. Systematic



3. Stratified



4. Cluster

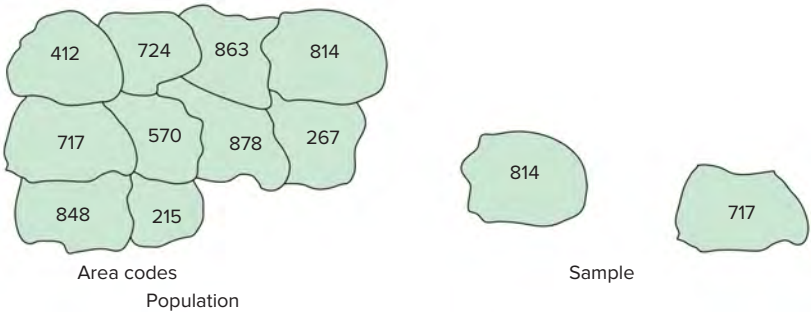


TABLE 1–3 Summary of Sampling Methods

Random	Subjects are selected by random numbers.
Systematic	Subjects are selected by using every k th number after the first subject is randomly selected from 1 through k .
Stratified	Subjects are selected by dividing up the population into subgroups (strata), and subjects are randomly selected within subgroups.
Cluster	Subjects are selected by using an intact subgroup that is representative of the population.

Since samples are not perfect representatives of the populations from which they are selected, there is always some error in the results. This error is called a *sampling error*.

Sampling error is the difference between the results obtained from a sample and the results obtained from the population from which the sample was selected.

For example, suppose you select a sample of full-time students at your college and find 56% are female. Then you go to the admissions office and get the genders of all full-time students that semester and find that 54% are female. The difference of 2% is said to be due to sampling error.

In most cases, this difference is unknown, but it can be estimated. This process is shown in Chapter 7.

There is another error that occurs in statistics called *nonsampling error*.

A **nonsampling error** occurs when the data are obtained erroneously or the sample is biased, i.e., nonrepresentative.

For example, data could be collected by using a defective scale. Each weight might be off by, say, 2 pounds. Also, recording errors can be made. Perhaps the researcher wrote an incorrect data value.

Caution and vigilance should be used when collecting data.

Other sampling techniques, such as *sequential sampling*, *double sampling*, and *multi-stage sampling*, are explained in Chapter 14, along with a more detailed explanation of the four basic sampling techniques.

EXAMPLE 1–5 Sampling Methods

State which sampling method was used.

- Out of 10 hospitals in a municipality, a researcher selects one and collects records for a 24-hour period on the types of emergencies that were treated there.
- A researcher divides a group of students according to gender, major field, and low, average, and high grade point average. Then she randomly selects six students from each group to answer questions in a survey.
- The subscribers to a magazine are numbered. Then a sample of these people is selected using random numbers.
- Every 10th bottle of Energized Soda is selected, and the amount of liquid in the bottle is measured. The purpose is to see if the machines that fill the bottles are working properly.

SOLUTION

- Cluster
- Stratified
- Random
- Systematic

Applying the Concepts 1-3

American Culture and Drug Abuse

Assume you are a member of a research team and have become increasingly concerned about drug use by professional sports players as one of several factors affecting people's attitudes toward drug use in general. You set up a plan and conduct a survey on how people believe the American culture (television, movies, magazines, and popular music) influences illegal drug use. Your survey consists of 2250 adults and adolescents from around the country. A consumer group petitions you for more information about your survey. Answer the following questions about your survey.

1. What type of survey did you use (phone, mail, or interview)?
2. What are the advantages and disadvantages of the surveying methods you did not use?
3. What type of scores did you use? Why?
4. Did you use a random method for deciding who would be in your sample?
5. Which of the methods (stratified, systematic, cluster, volunteer, or convenience) did you use?
6. Why was that method more appropriate for this type of data collection?
7. If a convenience sample were obtained consisting of only adolescents, how would the results of the study be affected?

See page 38 for the answers.

Exercises 1-3

1. Name five ways that data can be collected.
2. What is meant by sampling error and nonsampling error?
3. Why are random numbers used in sampling, and how are random numbers generated?
4. Name and define the four basic sampling methods.

For Exercises 5–10, define a population that may have been used and explain how the sample might have been selected.

5. *Time* magazine reported that 83% of people with household earnings over \$200,000 have a bachelor's degree.
6. *Time* magazine reported that 25% of the world's prisoners and prisons are in the United States.
7. A researcher found that the average size of a household in the United States was 2.54 people.
8. Adults aged 19–50 need 1000 milligrams of calcium per day. (*Source: Institute of Medicine Report*)
9. Taking statins raises the risk of developing diabetes. (*Source: Journal of American Medical Association and other sources*)

10. The average January 2012 temperature in Boston was 34.2°F. This was 5.2° higher than the normal January average temperature. (*Source: AccuWeather.com*)

For Exercises 11–16, identify the sampling method that was used.

11. To check the accuracy of a machine filling coffee cups, every fifth cup is selected, and the number of ounces of coffee is measured.
12. To determine how long people exercise, a researcher interviews 5 people selected from a yoga class, 5 people selected from a weight-lifting class, 5 people selected from an aerobics class, and 5 people from swimming classes.
13. In a large school district, a researcher numbers all the full-time teachers and then randomly selects 30 teachers to be interviewed.
14. In a medical research study, a researcher selects a hospital and interviews all the patients that day.
15. For 15 minutes, all customers entering a selected Wal-Mart store on a specific day are asked how many miles from the store they live.
16. Ten counties in Pennsylvania are randomly selected to determine the average county real estate tax that the residents pay.

1-4 Experimental Design

OBJECTIVE 6

Explain the difference between an observational and an experimental study.

Observational and Experimental Studies

There are several different ways to classify statistical studies. This section explains two types of studies: *observational studies* and *experimental studies*.

In an **observational study**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

For example, in August 2015 (*The Verge*) asked “Tons of people are buying Fitbits, but are they actually using them?” Fitbit is a manufacturer of step counting devices. Only 9.5 million registered users out of 19 million are active (50%). The past data showed an active rate of 60% (10.9 million registered, 6.5 million active). Endeavour Partners states that 33% of buyers of step devices continue to use them after 6 months. In this study, the researcher merely observed what had happened to the Fitbit owners over a period of time. There was no type of research intervention.

There are three main types of observational studies. When all the data are collected at one time, the study is called a *cross-sectional study*. When the data are collected using records obtained from the past, the study is called a *retrospective study*. Finally, if the data are collected over a period of time, say, past and present, the study is called a *longitudinal study*.

Observational studies have advantages and disadvantages. One advantage of an observational study is that it usually occurs in a natural setting. For example, researchers can observe people’s driving patterns on streets and highways in large cities. Another advantage of an observational study is that it can be done in situations where it would be unethical or downright dangerous to conduct an experiment. Using observational studies, researchers can study suicides, rapes, murders, etc. In addition, observational studies can be done using variables that cannot be manipulated by the researcher, such as drug users versus nondrug users and right-handedness versus left-handedness.

Observational studies have disadvantages, too. As mentioned previously, since the variables are not controlled by the researcher, a definite cause-and-effect situation cannot be shown since other factors may have had an effect on the results. Observational studies can be expensive and time-consuming. For example, if one wanted to study the habitat of lions in Africa, one would need a lot of time and money, and there would be a certain amount of danger involved. Finally, since the researcher may not be using his or her own measurements, the results could be subject to the inaccuracies of those who collected the data. For example, if the researchers were doing a study of events that occurred in the 1800s, they would have to rely on information and records obtained by others from a previous era. There is no way to ensure the accuracy of these records.

The other type of study is called an *experimental study*.

In an **experimental study**, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Interesting Fact

The safest day of the week for driving is Tuesday.

For example, a study conducted at Virginia Polytechnic Institute and presented in *Psychology Today* divided female undergraduate students into two groups and had the students perform as many sit-ups as possible in 90 seconds. The first group was told only to “Do your best,” while the second group was told to try to increase the actual number of sit-ups done each day by 10%. After four days, the subjects in the group who were given the vague instructions to “Do your best” averaged 43 sit-ups, while the group that was given the more specific instructions to increase the number of sit-ups by 10% averaged 56 sit-ups by the last day’s session. The conclusion then was that athletes who were given specific goals performed better than those who were not given specific goals.

This study is an example of a statistical experiment since the researchers intervened in the study by manipulating one of the variables, namely, the type of instructions given to each group.

In a true experimental study, the subjects should be assigned to groups randomly. Also, the treatments should be assigned to the groups at random. In the sit-up study, the article did not mention whether the subjects were randomly assigned to the groups.

Sometimes when random assignment is not possible, researchers use intact groups. These types of studies are done quite often in education where already intact groups are available in the form of existing classrooms. When these groups are used, the study is said to be a **quasi-experimental study**. The treatments, though, should be assigned at random. Most articles do not state whether random assignment of subjects was used.

Statistical studies usually include one or more *independent variables* and one *dependent variable*.

The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory variable**. The resultant variable is called the **dependent variable** or the **outcome variable**.

The outcome variable is the variable that is studied to see if it has changed significantly because of the manipulation of the independent variable. For example, in the sit-up study, the researchers gave the groups two different types of instructions, general and specific. Hence, the independent variable is the type of instruction. The dependent variable, then, is the resultant variable, that is, the number of sit-ups each group was able to perform after four days of exercise. If the differences in the dependent or outcome variable are large and other factors are equal, these differences can be attributed to the manipulation of the independent variable. In this case, specific instructions were shown to increase athletic performance.

In the sit-up study, there were two groups. The group that received the special instruction is called the **treatment group** while the other is called the **control group**. The treatment group receives a specific treatment (in this case, instructions for improvement) while the control group does not.

Both types of statistical studies have advantages and disadvantages. Experimental studies have the advantage that the researcher can decide how to select subjects and how to assign them to specific groups. The researcher can also control or manipulate the independent variable. For example, in studies that require the subjects to consume a certain amount of medicine each day, the researcher can determine the precise dosages and, if necessary, vary the dosage for the groups.

There are several disadvantages to experimental studies. First, they may occur in unnatural settings, such as laboratories and special classrooms. This can lead to several problems. One such problem is that the results might not apply to the natural setting. The age-old question then is, “This mouthwash may kill 10,000 germs in a test tube, but how many germs will it kill in my mouth?”

Another disadvantage with an experimental study is the **Hawthorne effect**. This effect was discovered in 1924 in a study of workers at the Hawthorne plant of the Western Electric Company. In this study, researchers found that the subjects who knew they were participating in an experiment actually changed their behavior in ways that affected the results of the study.

Another problem when conducting statistical studies is called *confounding of variables* or *lurking variables*.

A **confounding variable** is one that influences the dependent or outcome variable but was not separated from the independent variable.

Researchers try to control most variables in a study, but this is not possible in some studies. For example, subjects who are put on an exercise program might also improve their diet unbeknownst to the researcher and perhaps improve their health in other ways not due to exercise alone. Then diet becomes a confounding variable.

Interesting Fact

The number of potholes in the United States is about 56 million.

When you read the results of statistical studies, decide if the study was observational or experimental. Then see if the conclusion follows logically, based on the nature of these studies.

Another factor that can influence statistical experiments is called the *placebo effect*. Here the subjects used in the study respond favorably or show improvement due to the fact that they had been selected for the study. They could also be reacting to clues given unintentionally by the researchers. For example, in a study on knee pain done at the Houston VA Medical Center, researchers divided 180 patients into three groups. Two groups had surgery to remove damaged cartilage while those in the third group had simulated surgery. After two years, an equal number of patients in each group reported that they felt better after the surgery. Those patients who had simulated surgery were said to be responding to what is called the placebo effect.

To minimize the placebo effect, researchers use what is called *blinding*. In blinding, the subjects do not know whether they are receiving an actual treatment or a placebo. Many times researchers use a sugar pill that looks like a real medical pill. Often *double blinding* is used. Here both the subjects and the researchers are not told which groups are given the placebos.

Researchers use *blocking* to minimize variability when they suspect that there might be a difference between two or more blocks. For example, in the sit-up study mentioned earlier, if we think that men and women would respond differently to “Do your best” versus “Increase by 10% every day,” we would divide the subjects into two blocks (men, women) and then randomize which subjects in each block get the treatment.

When subjects are assigned to groups randomly, and the treatments are assigned randomly, the experiment is said to be a *completely randomized design*.

Some experiments use what is called a *matched-pair design*. Here one subject is assigned to a treatment group, and another subject is assigned to a control group. But, before the assignment, subjects are paired according to certain characteristics. In earlier years, studies used identical twins, assigning one twin to one group and the other twin to another group. Subjects can be paired on any characteristics such as ages, heights, and weights.

Another way to validate studies is to use *replication*. Here the same experiment is done in another part of the country or in another laboratory. The same study could also be done using adults who are not going to college instead of using college students. Then the results of the second study are compared to the ones in the original study to see if they are the same.

No matter what type of study is conducted, two studies on the same subject sometimes have conflicting conclusions. Why might this occur? An article titled “Bottom Line: Is It Good for You?” (*USA TODAY Weekend*) states that in the 1960s studies suggested that margarine was better for the heart than butter since margarine contains less saturated fat and users had lower cholesterol levels. In a 1980 study, researchers found that butter was better than margarine since margarine contained trans-fatty acids, which are worse for the heart than butter’s saturated fat. Then in a 1998 study, researchers found that margarine was better for a person’s health. Now, what is to be believed? Should one use butter or margarine?

The answer here is that you must take a closer look at these studies. Actually, it is not the choice between butter and margarine that counts, but the type of margarine used. In the 1980s, studies showed that solid margarine contains trans-fatty acids, and scientists believe that they are worse for the heart than butter’s saturated fat. In the 1998 study, liquid margarine was used. It is very low in trans-fatty acids, and hence it is more healthful than butter because trans-fatty acids have been shown to raise cholesterol. Hence, the conclusion is that it is better to use liquid margarine than solid margarine or butter.

Before decisions based on research studies are made, it is important to get all the facts and examine them in light of the particular situation.

The purpose of a statistical study is to gain and process information obtained from the study in order to answer specific questions about the subject being investigated. Statistical researchers use a specific procedure to do statistical studies to obtain valid results.

Unusual Stat

The chance that someone will attempt to burglarize your home in any given year is 1 in 20.

The general guidelines for this procedure are as follows:

1. Formulate the purpose of the study.
2. Identify the variables for the study.
3. Define the population.
4. Decide what sampling method you will use to collect the data.
5. Collect the data.
6. Summarize the data and perform any statistical calculations needed.
7. Interpret the results.

There is also a formal way to write up the study procedure and the results obtained. This information is available on the online resources under “Writing the Research Report.”

EXAMPLE 1–6 Experimental Design

Researchers randomly assigned 10 people to each of three different groups. Group 1 was instructed to write an essay about the hassles in their lives. Group 2 was instructed to write an essay about circumstances that made them feel thankful. Group 3 was asked to write an essay about events that they felt neutral about. After the exercise, they were given a questionnaire on their outlook on life. The researchers found that those who wrote about circumstances that made them feel thankful had a more optimistic outlook on life. The conclusion is that focusing on the positive makes you more optimistic about life in general. Based on this study, answer the following questions.

- a. Was this an observational or experimental study?
- b. What is the independent variable?
- c. What is the dependent variable?
- d. What may be a confounding variable in this study?
- e. What can you say about the sample size?
- f. Do you agree with the conclusion? Explain your answer.

SOLUTION

- a. This is an experimental study since the variables (types of essays written) were manipulated.
- b. The independent variable was the type of essay the participants wrote.
- c. The dependent variable was the score on the life outlook questionnaire.
- d. Other factors, such as age, upbringing, and income, can affect the results; however, the random assignment of subjects is helpful in eliminating these factors.
- e. In this study, the sample uses 30 participants total.
- f. Answers will vary.

OBJECTIVE 7

Explain how statistics can be used and misused.

Uses and Misuses of Statistics

As explained previously, statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypotheses, and make estimates about population characteristics. However, there is another aspect of statistics, and that is the misuse of statistical techniques to sell products that don't work properly, to attempt to prove something true that is really not true, or to get our attention by using statistics to evoke fear, shock, and outrage.

Two sayings that have been around for a long time illustrate this point:

“There are three types of lies—lies, damn lies, and statistics.”

“Figures don’t lie, but liars figure.”

Just because we read or hear the results of a research study or an opinion poll in the media, this does not mean that these results are reliable or that they can be applied to any and all situations. For example, reporters sometimes leave out critical details such as the size of the sample used or how the research subjects were selected. Without this information, you cannot properly evaluate the research and properly interpret the conclusions of the study or survey.

It is the purpose of this section to show some ways that statistics can be misused. You should not infer that all research studies and surveys are suspect, but that there are many factors to consider when making decisions based on the results of research studies and surveys. Here are some ways that statistics can be misrepresented.

Suspect Samples The first thing to consider is the sample that was used in the research study. Sometimes researchers use very small samples to obtain information. Several years ago, advertisements contained such statements as “Three out of four doctors surveyed recommend brand such and such.” If only 4 doctors were surveyed, the results could have been obtained by chance alone; however, if 100 doctors were surveyed, the results were probably not due to chance alone.

Not only is it important to have a sample size that is large enough, but also it is necessary to see how the subjects in the sample were selected. As stated previously, studies using volunteers sometimes have a built-in bias. Volunteers generally do not represent the population at large. Sometimes they are recruited from a particular socioeconomic background, and sometimes unemployed people volunteer for research studies to get a stipend. Studies that require the subjects to spend several days or weeks in an environment other than their home or workplace automatically exclude people who are employed and cannot take time away from work. Sometimes only college students or retirees are used in studies. In the past, many studies have used only men, but have attempted to generalize the results to both men and women. Opinion polls that require a person to phone or mail in a response most often are not representative of the population in general, since only those with strong feelings for or against the issue usually call or respond by mail.

Another type of sample that may not be representative is the convenience sample. Educational studies sometimes use students in intact classrooms since it is convenient. Quite often, the students in these classrooms do not represent the student population of the entire school district.

When results are interpreted from studies using small samples, convenience samples, or volunteer samples, care should be used in generalizing the results to the entire population.

Ambiguous Averages In Chapter 3, you will learn that there are four commonly used measures that are loosely called *averages*. They are the *mean*, *median*, *mode*, and *mid-range*. For the same data set, these averages can differ markedly. People who know this can, without lying, select the one measure of average that lends the most evidence to support their position.

Changing the Subject Another type of statistical distortion can occur when different values are used to represent the same data. For example, one political candidate who is running for reelection might say, “During my administration, expenditures increased a mere 3%.” His opponent, who is trying to unseat him, might say, “During my opponent’s administration, expenditures have increased a whopping \$6,000,000.” Here both figures are correct; however, expressing a 3% increase as \$6,000,000 makes it sound like a very large increase. Here again, ask yourself, Which measure better represents the data?

Detached Statistics A claim that uses a detached statistic is one in which no comparison is made. For example, you may hear a claim such as “Our brand of crackers has one-third fewer calories.” Here, no comparison is made. One-third fewer calories than what? Another example is a claim that uses a detached statistic such as “Brand A aspirin works four times faster.” Four times faster than what? When you see statements such as this, always ask yourself, Compared to what?

Implied Connections Many claims attempt to imply connections between variables that may not actually exist. For example, consider the following statement: “Eating fish may help to reduce your cholesterol.” Notice the words *may help*. There is no guarantee that eating fish will definitely help you reduce your cholesterol.

“Studies suggest that using our exercise machine will reduce your weight.” Here the word *suggest* is used; and again, there is no guarantee that you will lose weight by using the exercise machine advertised.

Another claim might say, “Taking calcium will lower blood pressure in some people.” Note the word *some* is used. You may not be included in the group of “some” people. Be careful when you draw conclusions from claims that use words such as *may*, *in some people*, and *might help*.

Misleading Graphs Statistical graphs give a visual representation of data that enables viewers to analyze and interpret data more easily than by simply looking at numbers. In Chapter 2, you will see how some graphs are used to represent data. However, if graphs are drawn inappropriately, they can misrepresent the data and lead the reader to draw false conclusions. The misuse of graphs is also explained in Chapter 2.

Faulty Survey Questions When analyzing the results of a survey using questionnaires, you should be sure that the questions are properly written since the way questions are phrased can often influence the way people answer them. For example, the responses to a question such as “Do you feel that the North Huntingdon School District should build a new football stadium?” might be answered differently than a question such as “Do you favor increasing school taxes so that the North Huntingdon School District can build a new football stadium?” Each question asks something a little different, and the responses could be radically different. When you read and interpret the results obtained from questionnaire surveys, watch out for some of these common mistakes made in the writing of the survey questions.

In Chapter 14, you will find some common ways that survey questions could be misinterpreted by those responding, and could therefore result in incorrect conclusions.

In summary then, statistics, when used properly, can be beneficial in obtaining much information, but when used improperly, can lead to much misinformation. It is like your automobile. If you use your automobile to get to school or work or to go on a vacation, that’s good. But if you use it to run over your neighbor’s dog because it barks all night long and tears up your flower garden, that’s not so good!

Applying the Concepts 1–4

Today’s Cigarettes

Vapor or electronic cigarettes have increased dramatically over the past five years. Compared to traditional tobacco products a lot of research has not been performed on alternative smoking devices even though the sales of these devices have double and then tripled. One study conducted at Virginia Commonwealth University considered the following factors: carbon monoxide concentration, heart rate, subjective effects of the user, and plasma nicotine concentration. The study consisted of 32 subjects. The subjects came from four separate groups, traditional cigarettes, 18-mg nicotine cartridge vapor brand, 16-mg nicotine cartridge vapor brand, and a device containing no vapors.

After five minutes of smoking, the smokers of the traditional cigarettes saw increased levels of carbon monoxide, heart rate, and plasma nicotine. The electronic vapor cigarette smokers did not see a significant increase in heart rate or subjective effects to the user. Answer the following questions.

1. What type of study was this (observational, quasi-experimental, or experimental)?
2. What are the independent and dependent variables?
3. Which was the treatment group?
4. Could the subjects' blood pressures be affected by knowing that they are part of a study?
5. List some possible confounding variables.
6. Do you think this is a good way to study the effect of smokeless tobacco?

See pages 38–39 for the answers.

Exercises 1–4

1. Explain the difference between an observational and an experimental study.
2. Name and define the three types of observational studies.
3. List some advantages and disadvantages of an observational study.
4. List some advantages and disadvantages of an experimental study.
5. What is the difference between an experimental study and a quasi-experimental study?
6. What is the difference between independent variables and dependent variables?
7. Why are a treatment group and a control group used in a statistical study?
8. Explain the Hawthorne effect.
9. What is a confounding variable?
10. Define the placebo effect in a statistical study.
11. What is meant by blinding and double-blinding?
12. Why do researchers use randomization in statistical studies?
13. What is the difference between a completely randomized design and a matched-pair design?
14. Why is replication used in statistical studies?

For Exercises 15–18, determine whether an observational study or an experimental study was used.

15. A survey was taken to see how many times in a month a person encountered a panhandler on the street.
16. A study using college students was conducted to see if the percentage of males who pay for a date was equal to

the percentage of females who pay for a date.

17. A study was done on two groups of overweight individuals. Group 1 was placed on a healthy, moderate diet. Group 2 was not given any diet instructions. After 1 month, the members were asked how many times they engaged in binge eating. The results of the two groups were compared.
18. Two groups of students were randomly selected. The students in Group 1 were enrolled in the general studies program. Group 2 students were enrolled in a specific major program (i.e., business, engineering, social work, criminal justice, etc.). At the end of the first year of study, the grade point averages of the two groups were compared.

In Exercises 19–22, identify the independent variable and the dependent variable.

19. According to the *British Journal of Sports Medicine*, a regular 30-minute workout could slash your risk of catching a cold by 43%.
20. The *Journal of Behavioral Medicine* reported that sharing a hug and holding hands can limit the physical effects of stress such as soaring heart rate and elevated blood pressure.
21. A study was conducted to determine whether when a restaurant server drew a happy face on the check, that would increase the amount of the tip.
22. A study was conducted to determine if the marital status of an individual had any effect on the cause of death of the individual.

For Exercises 23–26, suggest some confounding variables that the researcher might want to consider when doing a study.

23. *Psychology Today* magazine reports that the more intelligent a person is (based on IQ), the more willing the person is to make a cooperative choice rather than a selfish one.
24. *The New England Journal of Medicine* reported that when poor women move to better neighborhoods, they lower the risk of developing obesity and diabetes.
25. A leading journal reported that people who have a more flexible work schedule are more satisfied with their jobs.
26. York University in Toronto, Canada, stated that people who had suffered from fibromyalgia were able to reduce their pain by participating in twice-weekly yoga sessions.

For Exercises 27–31, give a reason why the statement made might be misleading.

27. Our product will give you the perfect body.
28. Here is the whole truth about back pain.
29. Our pain medicine will give you 24 hours of pain relief.
30. By reading this book, you will increase your IQ by 20 points.
31. Eating 21 grams of fiber may help you to lose weight.
32. List the steps you should perform when conducting a statistical study.
33. **Beneficial Bacteria** According to a pilot study of 20 people conducted at the University of Minnesota, daily doses of a compound called arabinogalactan over a period of 6 months resulted in a significant increase in the beneficial lactobacillus species of bacteria. Why can't it be concluded that the compound is beneficial for the majority of people?
34. Comment on the following statement, taken from a magazine advertisement: "In a recent clinical study, Brand ABC (actual brand will not be named) was proved to be 1950% better than creatine!"
35. In an ad for women, the following statement was made: "For every 100 women, 91 have taken the road less traveled." Comment on this statement.
36. In many ads for weight loss products, under the product claims and in small print, the following statement is made: "These results are not typical." What does this say about the product being advertised?
37. In an ad for moisturizing lotion, the following claim is made: "... it's the number 1 dermatologist-recommended brand." What is misleading about this claim?
38. An ad for an exercise product stated: "Using this product will burn 74% more calories." What is misleading about this statement?
39. "Vitamin E is a proven antioxidant and may help in fighting cancer and heart disease." Is there anything ambiguous about this claim? Explain.
40. "Just 1 capsule of Brand X can provide 24 hours of acid control." (Actual brand will not be named.) What needs to be more clearly defined in this statement?
41. "... Male children born to women who smoke during pregnancy run a risk of violent and criminal behavior that lasts well into adulthood." Can we infer that smoking during pregnancy is responsible for criminal behavior in people?
42. **Caffeine and Health** In the 1980s, a study linked coffee to a higher risk of heart disease and pancreatic cancer. In the early 1990s, studies showed that drinking coffee posed minimal health threats. However, in 1994, a study showed that pregnant women who drank 3 or more cups of tea daily may be at risk for miscarriage. In 1998, a study claimed that women who drank more than a half-cup of caffeinated tea every day may actually increase their fertility. In 1998, a study showed that over a lifetime, a few extra cups of coffee a day can raise blood pressure, heart rate, and stress (*Source*: "Bottom Line: Is It Good for You? Or Bad?" by Monika Guttman, *USA TODAY Weekend*). Suggest some reasons why these studies appear to be conflicting.

Extending the Concepts

43. Find an article that describes a statistical study, and identify the study as observational or experimental.
44. For the article that you used in Exercise 43, identify the independent variable(s) and dependent variable for the study.

45. For the article that you selected in Exercise 43, suggest some confounding variables that may have an effect on the results of the study.
46. Select a newspaper or magazine article that involves a statistical study, and write a paper answering these questions.
- Is this study descriptive or inferential? Explain your answer.
 - What are the variables used in the study? In your opinion, what level of measurement was used to obtain the data from the variables?
 - Does the article define the population? If so, how is it defined? If not, how could it be defined?
 - Does the article state the sample size and how the sample was obtained? If so, determine the size of the sample and explain how it was selected. If not, suggest a way it could have been obtained.
 - Explain *in your own words* what procedure (survey, comparison of groups, etc.) might have been used to determine the study's conclusions.
 - Do you agree or disagree with the conclusions? State your reasons.

1-5 Computers and Calculators

OBJECTIVE 8

Explain the importance of computers and calculators in statistics.

In the past, statistical calculations were done with pencil and paper. However, with the advent of calculators, numerical computations became much easier. Computers do all the numerical calculation. All one does is to enter the data into the computer and use the appropriate command; the computer will print the answer or display it on the screen. Now the TI-84 Plus graphing calculator accomplishes the same thing.

There are many statistical packages available. This book uses Microsoft Excel and MINITAB. Instructions for using the TI-84 Plus graphing calculator, Excel, and MINITAB have been placed at the end of each relevant section, in subsections entitled Technology Step by Step.

You should realize that the computer and calculator merely give numerical answers and save the time and effort of doing calculations by hand. You are still responsible for understanding and interpreting each statistical concept. In addition, you should realize that the results come from the data and do not appear magically on the computer. Doing calculations by using the procedure tables will help you reinforce this idea.

The author has left it up to instructors to choose how much technology they will incorporate into the course.

Technology

TI-84 Plus Step by Step

Step by Step

The TI-84 Plus graphing calculator can be used for a variety of statistical graphs and tests.

General Information

To turn calculator on:
Press **ON** key.

To turn calculator off:
Press **2nd** [**OFF**].

To reset defaults only:

- Press **2nd**, then [**MEM**].
- Select **7**, then **2**, then **2**.

Optional. To reset settings on calculator and clear memory (*note*: this will clear all settings and programs in the calculator's memory):

Press **2nd**, then [**MEM**]. Then press **7**, then **1**, then **2**.

(Also, the contrast may need to be adjusted after this.)

To adjust contrast (if necessary):

Press **2nd**. Then press and hold **▲** to darken or **▼** to lighten contrast.

To clear screen:

Press **CLEAR**.

(*Note:* This will return you to the screen you were using.)

To display a menu:

Press appropriate menu key. Example: **STAT**.

To return to home screen:

Press **2nd**, then **[QUIT]**.

To move around on the screens:

Use the arrow keys.

To select items on the menu:

Press the corresponding number or move the cursor to the item, using the arrow keys. Then press **ENTER**.

(*Note:* In some cases, you do not have to press **ENTER**, and in other cases you may need to press **ENTER** twice.)

Entering Data

To enter single-variable data (clear the old list if necessary, see “Editing Data”):

1. Press **STAT** to display the Edit menu.
2. Press **ENTER** to select 1:Edit.
3. Enter the data in L_1 and press **ENTER** after each value.
4. After all data values are entered, press **STAT** to get back to the Edit menu or **2nd [QUIT]** to end.

Example TI1–1

Enter the following data values in L_1 : **213, 208, 203, 215, 222**.

To enter multiple-variable data:

The TI-84 Plus will take up to six lists designated

L_1 , L_2 , L_3 , L_4 , L_5 , and L_6 .

1. To enter more than one set of data values, complete the preceding steps. Then move the cursor to L_2 by pressing the **►** key.
2. Repeat the steps in the preceding part.

Output

L_1	L_2	L_3	1
213	-----	-----	
208			
203			
215			
222			

$L_1(6)=$			

Editing Data

To correct a data value before pressing **ENTER**, use **◀** and retype the value and press **ENTER**.

To correct a data value in a list after pressing **ENTER**, move the cursor to the incorrect value in list and type in the correct value. Then press **ENTER**.

To delete a data value in a list:

Move the cursor to a value and press **DEL**.

To insert a data value in a list:

1. Move cursor to position where data value is to be inserted; then press **2nd [INS]**.
2. Type data value; then press **ENTER**.

To clear a list:

1. Press **STAT**, then **4**.
2. Enter list to be cleared. Example: To clear L_1 , press **2nd [L_1]**. Then press **ENTER**.

(*Note:* To clear several lists, follow Step 1, but enter each list to be cleared, separating them with commas. To clear all lists at once, follow Step 1; then press **ENTER**.)

Sorting Data

To sort the data in a list:

1. Enter the data in L_1 .
2. Press **STAT**, and then **2** to get **SortA(** to sort the list in ascending order.
3. Then press **2nd [L_1]** **ENTER**.

The calculator will display Done.

4. Press **STAT** and then **ENTER** to display the sorted list.
(Note: The SortD(or 3 sorts the list in descending order.)

Example TI1-2

Sort in ascending order the data values entered in Example TI1-1.

Output

L1	L2	L3	1
203	-----	-----	
208			
213			
215			
215			
222			
222			
L1(6)=			

EXCEL Step by Step

Excel's Analysis
ToolPak Add-In

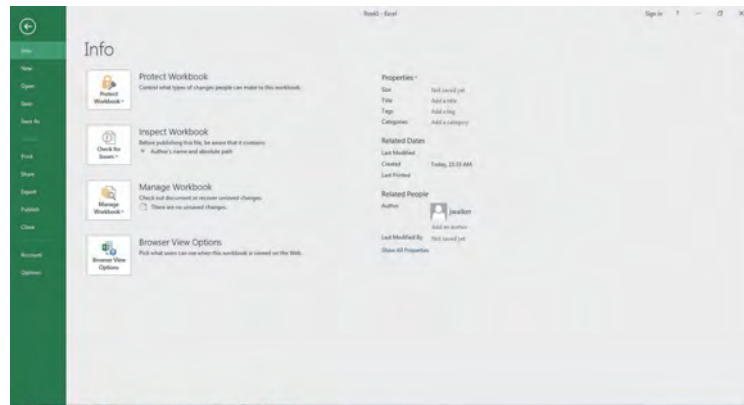
General Information

Microsoft Excel 2010 has two different ways to solve statistical problems. First, there are built-in functions, such as STDEV and CHITEST, available from the standard toolbar by

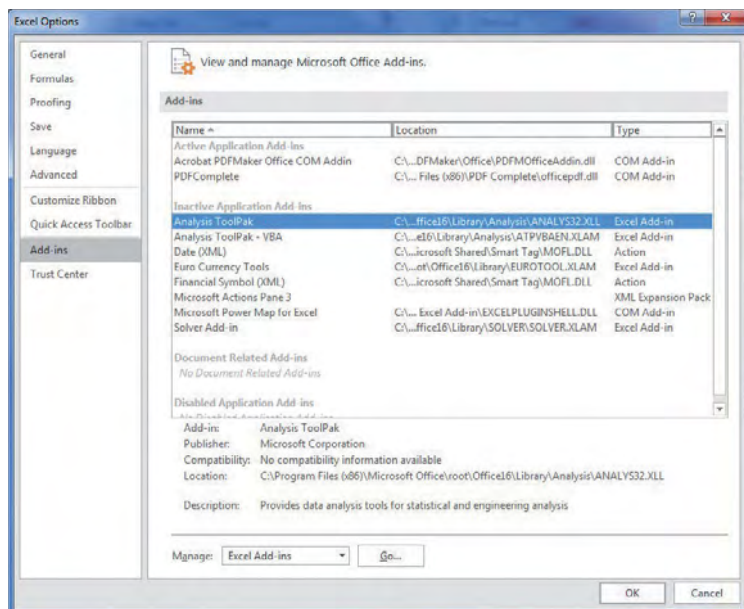
clicking Formulas, and then selecting the Insert Function icon . Another feature of Excel that is useful for calculating multiple statistical measures and performing statistical tests for a set of data is the Data Analysis command found in the Analysis ToolPak Add-in.

To load the Analysis ToolPak:

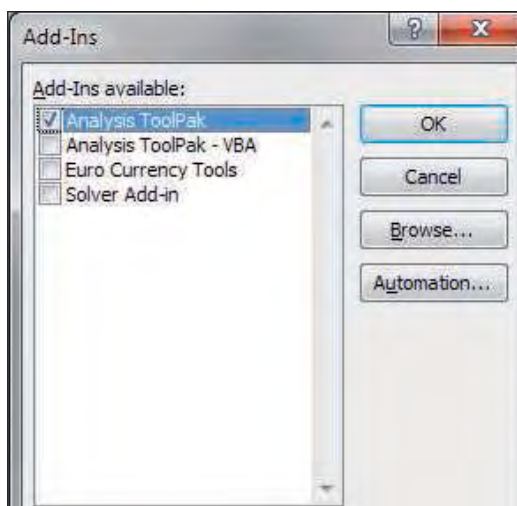
Click the File tab in the upper left-hand corner of an Excel workbook, then select Options in the left-hand panel.



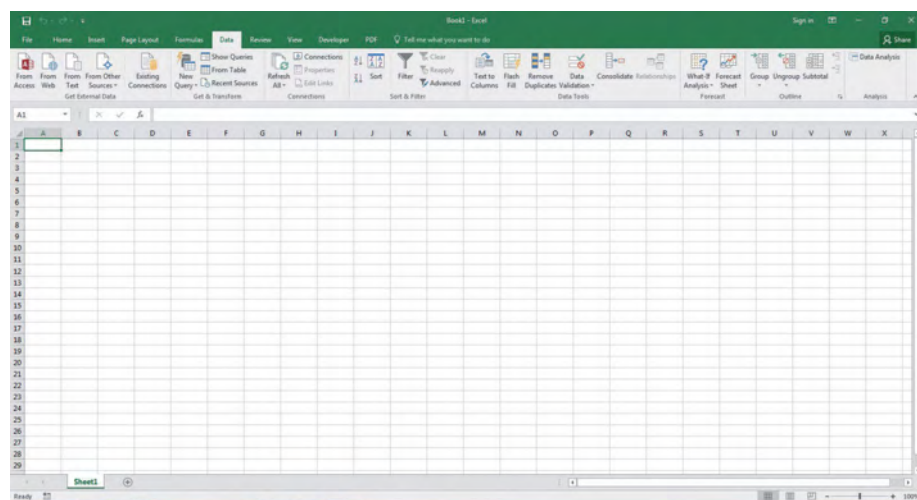
1. Click Add-Ins, and then click the Go button at the bottom of the Excel Options page to the right of the Manage tool.



2. Check the Analysis ToolPak Add-in and click OK.



3. Click the Data Tab. The Analysis ToolPak will appear in the Analysis group.

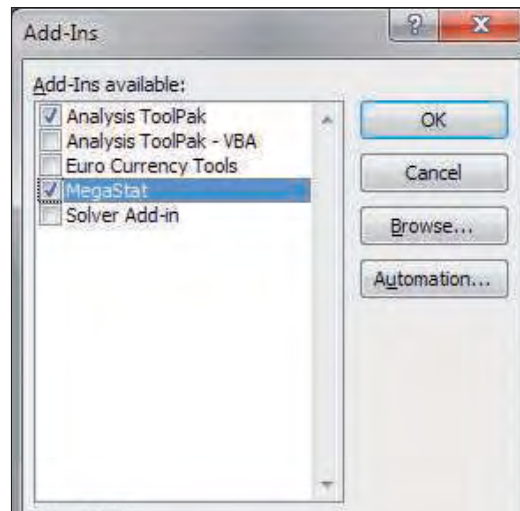


MegaStat

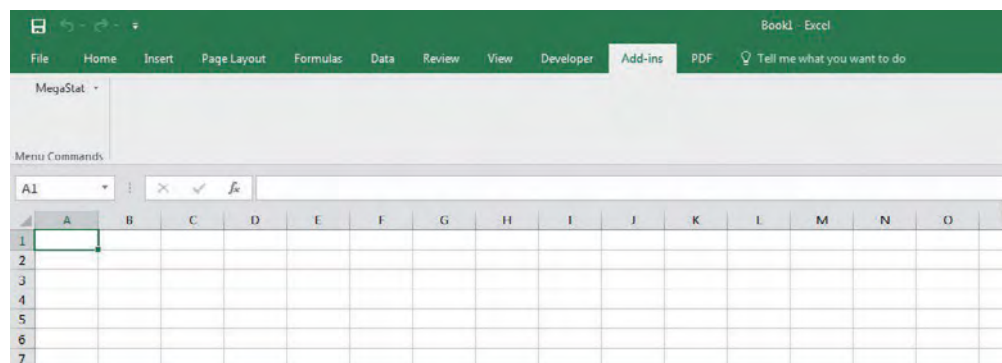
Later in this text you will encounter a few Excel Technology Step by Step operations that will require the use of the MegaStat Add-in for Excel. MegaStat can be purchased from www.mhhe.com/megastat.

1. Save the Zip file containing the MegaStat Excel Add-in file (MegaStat.xla) and the associated help file on your computer's hard drive.
2. Open the Excel software.
3. Click the File tab and select Options (as before with the installation of the Analysis ToolPak).
4. Click the Add-Ins button. MegaStat will not appear as an Application until first installation.
5. Click Go button next to the Manage (Add-Ins) Tool at the bottom of the Excel Options window.
6. Once the Add-Ins Checkbox appears, click Browse to locate the MegaStat.xla file on your computer's hard drive.

7. Select the MegaStat.xla file from the hard drive; click the Checkbox next to it and click OK.



8. The MegaStat Add-in will appear when you select the Add-Ins tab in the Toolbar.



Entering Data

1. Select a cell at the top of a column on an Excel worksheet where you want to enter data.
When working with data values for a single variable, you will usually want to enter the values into a single column.
2. Type each data value and press **[Enter]** or **[Tab]** on your keyboard.

You can also add more worksheets to an Excel workbook by clicking the Insert Worksheet icon  located at the bottom of an open workbook.

Example XL1-1: Opening an existing Excel workbook/worksheet

1. Open the Microsoft Office Excel 2010 program.
2. Click the File tab, then click Open.
3. Click the name of the library that contains the file, such as My documents, and then click Open.
4. Click the name of the Excel file that you want to open and then click Open.

Note: Excel files have the extension .xls or .xlsx.

MINITAB

Step by Step

General Information

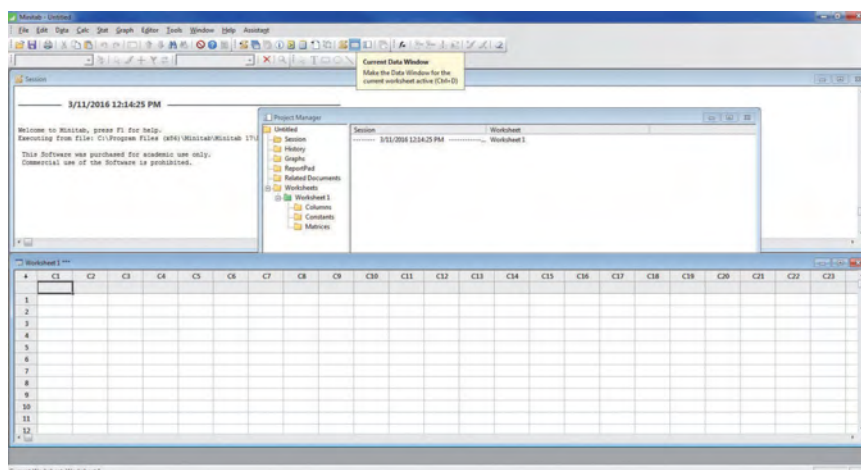
MINITAB statistical software provides a wide range of statistical analysis and graphing capabilities.

Take Note

In this text you will see captured MINITAB images from Windows computers running MINITAB Release 17. If you are using an earlier or later release of MINITAB, the screens you see on your computer may bear slight visual differences from the screens pictured in this text.

Start the Program

1. Click the Windows Start Menu, then All Programs.
2. Click the MINITAB folder and then click  the program icon. The program screen will look similar to the one shown here. You will see the Session Window, the Worksheet Window, and perhaps the Project Manager Window.
3. Click the Project Manager icon on the toolbar to bring the project manager to the front.



To use the program, data must be entered from the keyboard or from a file.

Entering Data from the Keyboard

In MINITAB, all the data for one variable are stored in a column. Step by step instructions for entering these data follow.

Data

213 208 203 215 222

1. Click in row 1 of Worksheet 1***. This makes the worksheet the active window and puts the cursor in the first cell. The small data entry arrow in the upper left-hand corner of the worksheet should be pointing down. If it is not, click it to change the direction in which the cursor will move when you press the [Enter] key.
2. Type in each number, pressing [Enter] after each entry, including the last number typed.
3. *Optional:* Click in the space above row 1 to type in **Weight**, the column label.

Worksheet 1	
	C1
	Weight
1	213
2	208
3	203
4	215
5	222

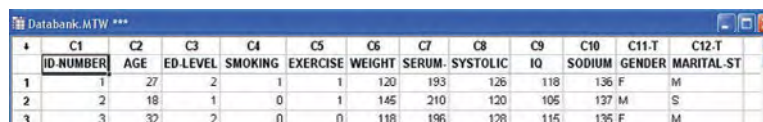
Save a Worksheet File

4. Click on the **File Menu**. *Note:* This is *not* the same as clicking the disk icon .
5. Click **Save Current Worksheet As . . .**
6. In the dialog box you will need to verify three items:
 - a) Save in: Click on or type in the disk drive and directory where you will store your data. This may be a thumb drive such as E:\ or a hard-drive folder such as C:\MinitabData.
 - b) File Name: Type in the name of the file, such as **MyData**.
 - c) Save as Type: The default here is MINITAB. An extension of mtw is added to the name. Click [Save]. The name of the worksheet will change from Worksheet 1*** to MyData.MTW***. The triple asterisks indicate the active worksheet.

Open the Databank File

The raw data are shown in Appendix B. There is a row for each person's data and a column for each variable. MINITAB data files comprised of data sets used in this book, including the Databank, are available at www.mhhe.com/bluman. Here is how to get the data from a file into a worksheet.

1. Click **File>Open Worksheet**. A sequence of menu instructions will be shown this way.
Note: This is *not* the same as clicking the file icon . If the dialog box says Open Project instead of Open Worksheet, click [Cancel] and use the correct menu item. The Open Worksheet dialog box will be displayed.
2. You must check three items in this dialog box.
 - a) The Look In: dialog box should show the directory where the file is located.
 - b) Make sure the Files of Type: shows the correct type, MINITAB [*.mtw].
 - c) Double-click the file name in the list box Databank.mtw. A dialog box may inform you that a copy of this file is about to be added to the project. Click on the checkbox if you do not want to see this warning again.
3. Click the [OK] button. The data will be copied into a second worksheet. Part of the worksheet is shown here.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11-T	C12-T
	ID NUMBER	AGE	ED LEVEL	SMOKING	EXERCISE	WEIGHT	SERUM	SYSTOLIC	IQ	SODIUM	GENDER	MARITAL ST
1	1	27	2	1	1	120	193	126	118	136	F	M
2	2	18	1	0	1	145	210	120	105	137	M	S
3	3	32	2	0	0	118	196	128	115	135	F	M

- a) You may maximize the window and scroll if desired.
 - b) C12-T Marital Status has a T appended to the label to indicate alphanumeric data. MyData.MTW is not erased or overwritten. Multiple worksheets can be available; however, only the active worksheet is available for analysis.
4. To switch between the worksheets, select **Window>MyData.MTW**.
5. Select **File>Exit** to quit. To save the project, click [Yes].
6. Type in the name of the file, **Chapter01**. The Data Window, the Session Window, and settings are all in one file called a project. Projects have an extension of mpj instead of mtw.

Clicking the disk icon  on the menu bar is the same as selecting **File>Save Project**.

Clicking the file icon  is the same as selecting **File>Open Project**.
7. Click [Save]. The mpj extension will be added to the name. The computer will return to the Windows desktop. The two worksheets, the Session Window results, and settings are saved in this project file. When a project file is opened, the program will start up right where you left off.

Summary

- The two major areas of statistics are descriptive and inferential. Descriptive statistics includes the collection, organization, summarization, and presentation of data. Inferential statistics includes making inferences from samples to populations, estimations and hypothesis testing, determining relationships, and making predictions. Inferential statistics is based on *probability theory*. (1–1)*
- Data can be classified as qualitative or quantitative. Quantitative data can be either discrete or continuous, depending on the values they can assume. Data can also be measured by various scales. The four basic levels of measurement are nominal, ordinal, interval, and ratio. (1–2)
- Since in most cases the populations under study are large, statisticians use subgroups called samples to get the necessary data for their studies. There are four basic methods used to obtain samples: random, systematic, stratified, and cluster. (1–3)
- There are two basic types of statistical studies: observational studies and experimental studies. When conducting observational studies, researchers observe what is happening or what has happened and then draw conclusions based on these observations. They do not attempt to manipulate the variables in any way. (1–4)
- When conducting an experimental study, researchers manipulate one or more of the independent or explanatory variables and see how this manipulation influences the dependent or outcome variable. (1–4)
- Finally, the applications of statistics are many and varied. People encounter them in everyday life, such as in reading newspapers or magazines, listening to an MP3 player, or watching television. Since statistics is used in almost every field of endeavor, the educated individual should be knowledgeable about the vocabulary, concepts, and procedures of statistics. Also, everyone should be aware that statistics can be misused. (1–4)
- Today, computers and calculators are used extensively in statistics to facilitate the computations. (1–5)

*The numbers in parentheses indicate the chapter section where the material is explained.

Important Terms

blinding 20	dependent variable 19	measurement scales 8	random variable 3
blocking 20	descriptive statistics 3	nominal level of measurement 8	ratio level of measurement 8
boundary 7	discrete variables 6	nonsampling error 16	replication 20
census 3	double blinding 20	observational study 18	retrospective study 18
cluster sample 14	experimental study 18	ordinal level of measurement 8	sample 3
completely randomized design 20	explanatory variable 19	outcome variable 19	sampling error 16
confounding variable 19	Hawthorne effect 19	placebo effect 20	statistics 2
continuous variables 6	hypothesis testing 4	population 3	stratified sample 13
control group 19	independent variable 19	probability 4	systematic sample 12
convenience sample 14	inferential statistics 4	qualitative variables 6	treatment group 19
cross-sectional study 18	interval level of measurement 8	quantitative variables 6	variable 3
data 3	longitudinal study 18	quasi-experimental study 19	volunteer sample 14
data set 3	lurking variable 19	random sample 12	
data value or datum 3	matched-pair design 20		

Review Exercises

Section 1–1

For Exercises 1–8, state whether descriptive or inferential statistics has been used.

1. By 2040 at least 3.5 billion people will run short of water (World Future Society).
2. In a sample of 100 individuals, 36% think that watching television is the best way to spend an evening.
3. In a survey of 1000 adults, 34% said that they posted notes on social media websites (*Source: AARP Survey*).
4. In a poll of 3036 adults, 32% said that they got a flu shot at a retail clinic (*Source: Harris Interactive Poll*).
5. Allergy therapy makes bees go away (*Source: Prevention*).
6. Drinking decaffeinated coffee can raise cholesterol levels by 7% (*Source: American Heart Association*).
7. In a survey of 1500 people who gave up driving, the average of the ages at which they quit driving was 85. (*Men's Health*)
8. Experts say that mortgage rates may soon hit bottom (*Source: USA TODAY*).

Section 1–2

For Exercises 9–18, classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

9. Pages in the 25 best-selling mystery novels.
10. Rankings of golfers in a tournament.
11. Temperatures of 10 toasters.
12. Weights of selected cell phones.
13. Salaries of the coaches in the NFL.
14. Times required to complete a 6-mile bike ride.
15. Ratings of textbooks (poor, fair, good, excellent).
16. Number of amps delivered by battery chargers.
17. Ages of the players on a professional football team.
18. Categories of magazines in a physician's office (sports, women's, health, men's, news).

For Exercises 19–26, classify each variable as qualitative or quantitative.

19. Marital status of nurses in a hospital.
20. Time it takes 10 people to complete a *New York Times* crossword puzzle.
21. Weights of lobsters in a tank in a restaurant.
22. Colors of automobiles in a shopping center parking lot.
23. Amount of garbage (in pounds) discarded by residents of a high-rise apartment complex.

24. Capacity of the NFL football stadiums.
25. Ages of people living in a personal care home.
26. The different species of fish sold by a pet shop store.

For Exercises 27–34, classify each variable as discrete or continuous.

27. Number of street corner mailboxes in the city of Philadelphia.
28. Relative humidity levels in operating rooms at local hospitals.
29. Number of bananas in a bunch at several local supermarkets.
30. Ages of women when they were first married.
31. Weights of the backpacks of first-graders on a school bus.
32. Number of students each day who make appointments with a math tutor at a local college.
33. Duration of marriages in America (in years).
34. Ages of children in a preschool.

For Exercises 35–38, give the boundaries of each value.

35. 56 yards.
36. 105.4 miles.
37. 72.6 tons.
38. 9.54 millimeters.

Section 1–3

For Exercises 39–44, classify each sample as random, systematic, stratified, cluster, or other.

39. In a large school district, all teachers from two buildings are interviewed to determine whether they believe the students have less homework to do now than in previous years.
40. All fast-food workers at a randomly selected fast-food restaurant are selected and asked how many hours per week they work.
41. A group of unmarried men are selected using random numbers and asked how long it has been since their last date.
42. Every 100th hamburger manufactured is checked to determine its fat content.
43. Mail carriers of a large city are divided into four groups according to gender (male or female) and according to whether they walk or ride on their routes. Then 10 are selected from each group and interviewed to determine whether they have been bitten by a dog in the last year.
44. People are asked to phone in their response to a survey question.

Section 1–4

For Exercises 45–48, identify each study as being either observational or experimental.

45. Subjects were randomly assigned to two groups, and one group was given an herb and the other group a placebo. After 6 months, the numbers of respiratory tract infections each group had were compared.
46. A researcher stood at a busy intersection to see if the color of the automobile that a person drives is related to running red lights.
47. A sample of females were asked if their supervisors (bosses) at work were to hug them, would they consider that a form of sexual harassment?
48. Three groups of gamblers were randomly selected. The first group was given \$25 in casino money. The second group was given a \$25 coupon for food. The third group was given nothing. After a trip to the casino, each group was surveyed and asked their opinion of their casino experience.

For Exercises 49–52, identify the independent and dependent variables for each study.

49. A study was conducted to determine if crocodiles raised in captivity (i.e., in a zoo) grew faster than crocodiles living in the wild. Identify the explanatory variable and the outcome variable.
50. People who walk at least 3 miles a day are randomly selected, and their blood triglyceride levels are measured in order to determine if the number of miles that they walk has any influence on these levels.
51. In an article in the *British Journal of Nutrition*, two types of mice were randomly selected. One group received a thyme supplement for a specific time, while another group was used as a control group and received no supplements. The brains of the mice were then analyzed, and it was found that the brains of the group of mice that received the thyme supplements had antioxidant levels similar to those of younger mice. It was concluded that the thyme supplement increased the antioxidants in the brains of the mice.
52. A study was conducted to determine if workers who had a flexible work schedule had greater job satisfaction than those workers who worked a regular nine-to-five work schedule.

For Exercises 53–58, explain why the claims of these studies might be suspect.

53. Based on a recent telephone survey, 72% of those contacted shop online.
54. In High Point County there are 672 raccoons.
55. A survey of a group of people said the thing they dislike most about winter is snow.
56. Only 5% of men surveyed said that they liked “chick flicks.”
57. A recent study shows that high school dropouts spend less time on the Internet than those who graduated; therefore, the Internet raises your IQ.
58. Most shark attacks occur in ocean water that is 3 feet deep; therefore, it is safer to swim in deep water.

STATISTICS TODAY

Is Higher Education “Going Digital”? —Revisited

Researchers at the Pew Research Center used a telephone survey of 2142 graduates and an online survey of 1055 college and university presidents of two-year and four-year public and private colleges to ascertain their findings.

They found out that approximately 89% of public colleges and universities offer online classes while 60% of four-year private colleges offer them. About 23% of the graduates said that they have taken an online course. The college presidents predict that in 10 years, most of their students will have taken an online course.

As to the value of the online courses, 51% of the college presidents say that online courses are of equal value to classroom courses, but only 29% of the graduates say that they are of equal value.

Fifty-five percent of the college presidents said that plagiarism has increased over the last 10 years, and 89% said that the use of computers and the Internet have contributed to the increases.

Fifty-seven percent of recent college graduates said that they have used a laptop, smartphone, or computer tablet in the classroom. Most colleges have no rules for their use, but leave it up to the individual instructor to determine the limitations of their use.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

- Probability is used as a basis for inferential statistics.
- When the sample does not represent the population, it is called a biased sample.
- The difference between a sampling measure and a population measure is called a nonsampling error.
- When the population of college professors is divided into groups according to their rank (instructor, assistant professor, etc.) and then several are selected from each group to make up a sample, the sample is called a cluster sample.
- The variable temperature is an example of a quantitative variable.
- The height of basketball players is considered a continuous variable.
- The boundary of a value such as 6 inches would be 5.9–6.1 inches.

Select the best answer.

- The number of ads on a one-hour television show is what type of data?
 - Nominal
 - Qualitative
 - Discrete
 - Continuous
- What are the boundaries of 25.6 ounces?
 - 25–26 ounces
 - 25.55–25.65 ounces
 - 25.5–25.7 ounces
 - 20–39 ounces
- A researcher divided subjects into two groups according to gender and then selected members from each group for her sample. What sampling method was the researcher using?
 - Cluster
 - Random
 - Systematic
 - Stratified
- Data that can be classified according to color are measured on what scale?
 - Nominal
 - Ratio
 - Ordinal
 - Interval

- A study that involves no researcher intervention is called
 - An experimental study.
 - A noninvolvement study.
 - An observational study.
 - A quasi-experimental study.
- A variable that interferes with other variables in the study is called
 - A confounding variable.
 - An explanatory variable.
 - An outcome variable.
 - An interfering variable.

Use the best answer to complete these statements.

- Two major branches of statistics are _____ and _____.
- Two uses of probability are _____ and _____.
- The group of all subjects under study is called a(n) _____.
- A group of subjects selected from the group of all subjects under study is called a(n) _____.
- Three reasons why samples are used in statistics:
 - _____
 - _____
 - _____
- The four basic sampling methods are
 - _____
 - _____
 - _____
 - _____
- A study that uses intact groups when it is not possible to randomly assign participants to the groups is called a(n) _____ study.
- In a research study, participants should be assigned to groups using _____ methods, if possible.
- For each statement, decide whether descriptive or inferential statistics is used.
 - The average life expectancy in New Zealand is 78.49 years (*Source: World Factbook*).
 - A diet high in fruits and vegetables will lower blood pressure (*Source: Institute of Medicine*).
 - The total amount of estimated losses for Hurricane Katrina was \$125 billion (*Source: The World Almanac and Book of Facts*).
 - Researchers stated that the shape of a person's ears is relative to the person's aggression (*Source: American Journal of Human Biology*).

- e. In 2050, it is estimated that there will be 18 million Americans who are age 85 and over (*Source*: U.S. Census Bureau).
23. Classify each as nominal-level, ordinal-level, interval-level, or ratio-level of measurement.
- Rating of movies as G, PG, and R
 - Number of candy bars sold on a fund drive
 - Classification of automobiles as subcompact, compact, standard, and luxury
 - Temperatures of hair dryers
 - Weights of suitcases on a commercial airliner
24. Classify each variable as discrete or continuous.
- Ages of people working in a large factory
 - Number of cups of coffee served at a restaurant
 - The amount of drug injected into a guinea pig
 - The time it takes a student to drive to school
 - The number of gallons of milk sold each day at a grocery store
25. Give the boundaries of each.
- 32 minutes
 - 0.48 millimeter
 - 6.2 inches
 - 19 pounds
 - 12.1 quarts

Critical Thinking Challenges

- World's Busiest Airports** A study of the world's busiest airports was conducted by *Airports Council International*. Describe three variables that one could use to determine which airports are the busiest. What *units* would one use to measure these variables? Are these variables categorical, discrete, or continuous?
- Smoking and Criminal Behavior** The results of a study published in *Archives of General Psychiatry* stated that male children born to women who smoke during pregnancy run a risk of violent and criminal behavior that lasts into adulthood. The results of this study were challenged by some people in the media. Give several reasons why the results of this study would be challenged.
- Piano Lessons Improve Math Ability** The results of a study published in *Neurological Research* stated that second-graders who took piano lessons and played a computer math game more readily grasped math problems in fractions and proportions than a similar group who took an English class and played the same math game. What type of inferential study was this? Give several reasons why the piano lessons could improve a student's math ability.
- ACL Tears in Collegiate Soccer Players** A study of 2958 collegiate soccer players showed that in 46 anterior cruciate ligament (ACL) tears, 36 were in women. Calculate the percentages of tears for each gender.
 - Can it be concluded that female athletes tear their knees more often than male athletes?
 - Comment on how this study's conclusion might have been reached.

Data Projects

- Business and Finance** Investigate the types of data that are collected regarding stock and bonds, for example, price, earnings ratios, and bond ratings. Find as many types of data as possible. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.
- Sports and Leisure** Select a professional sport. Investigate the types of data that are collected about that sport, for example, in baseball, the level of play (A, AA, AAA, Major League), batting average, and home-runs. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.
- Technology** Music organization programs on computers and music players maintain information about a song, such as the writer, song length, genre, and your personal rating. Investigate the types of data collected about a song. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.

4. **Health and Wellness** Think about the types of data that can be collected about your health and wellness, things such as blood type, cholesterol level, smoking status, and body mass index. Find as many data items as you can. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.
5. **Politics and Economics** Every 10 years since 1790, the federal government has conducted a census of U.S. residents. Investigate the types of data that were collected in the 2010 census. For each, identify the level of measurement as nominal, ordinal, interval, or ratio.

For any quantitative data, also note if they are discrete or continuous. Use the library or a genealogy website to find a census form from 1860. What types of data were collected? How do the types of data differ?

6. **Your Class** Your school probably has a database that contains information about each student, such as age, county of residence, credits earned, and ethnicity. Investigate the types of student data that your college collects and reports. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.

Answers to Applying the Concepts

Section 1–1 Attendance and Grades

1. The variables are grades and attendance.
2. The data consist of specific grades and attendance numbers.
3. These are descriptive statistics; however, if an inference were made to all students, then that would be inferential statistics.
4. The population under study is all students at Manatee Community College (MCC).
5. While not specified, we probably have data from a sample of MCC students.
6. Based on the data, it appears that, in general, the better your attendance, the higher your grade.

Section 1–2 Fatal Transportation Injuries

1. The variables are transportation industry and fatal accidents.
2. Transportation industry is a qualitative variable, and the number of fatal accidents is a quantitative variable.
3. The number of fatalities is discrete.
4. The type of industry is nominal, and the number of fatalities is ratio.
5. Even though the number of fatalities for the railroad industry is lowest, you should consider the fact that fewer people use the railroads to travel than the other industries.
6. A person's transportation choice might also be affected by convenience, cost, service, availability, etc.
7. Answers will vary. The railroad industry had the fewest fatalities followed by water vehicle accidents while the aircraft accidents were about three times as many as the water vehicle accidents. Of course, the most fatalities occurred in highway accidents.

Section 1–3 American Culture and Drug Abuse

Answers will vary, so this is one possible answer.

1. I used a telephone survey. The advantage to my survey method is that this was a relatively inexpensive survey method (although more expensive than using the mail) that could get a fairly sizable response. The disadvantage to my survey method is that I have not included anyone without a telephone. (*Note:* My survey used a random dialing method to include unlisted numbers and cell phone exchanges.)
2. A mail survey also would have been fairly inexpensive, but my response rate may have been much lower than what I got with my telephone survey. Interviewing would have allowed me to use follow-up questions and to clarify any questions of the respondents at the time of the interview. However, interviewing is very labor- and cost-intensive.
3. I used ordinal data on a scale of 1 to 5. The scores were 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.
4. The random method that I used was a random dialing method.
5. To include people from each state, I used a stratified random sample, collecting data randomly from each of the area codes and telephone exchanges available.
6. This method allowed me to make sure that I had representation from each area of the United States.
7. Convenience samples may not be representative of the population, and a convenience sample of adolescents would probably differ greatly from the general population with regard to the influence of American culture on illegal drug use.

Section 1–4 Today's Cigarettes

1. This was an experiment, since the researchers imposed a treatment on each of the four groups involved in the study.

2. The independent variable was the smoking device. The dependent variables were carbon monoxide, heart rate, and plasma nicotine.
3. The treatment group was the group containing no vapors.
4. A subjects blood pressure might not be affected by knowing that he or she was part of a study. However, if the subject's blood pressure was affected by this knowledge, all the subjects (in all the groups) would be affected similarly. This might be an example of the Hawthorne effect.
5. Answers will vary. The age of the subjects, gender, previous smoking habits, and their physical fitness could be confounding variables.
6. Answers will vary. One possible answer is that the study design was fine, but that it cannot be generalized beyond the population of backgrounds of the subjects in the study.

Frequency Distributions and Graphs

STATISTICS TODAY

How Your Identity Can Be Stolen

Identity fraud is a big business today—more than 12.7 million people were victims. The total amount of the fraud in 2014 was \$16 billion. The average amount of the fraud for a victim is \$1260, and the average time to correct the problem is 40 hours. The ways in which a person's identity can be stolen are presented in the following table:

Government documents or benefits fraud	38.7%
Credit card fraud	17.4
Phone or utilities fraud	12.5
Bank fraud	8.2
Attempted identity theft	4.8
Employment-related fraud	4.8
Loan fraud	4.4
Other identity theft	9.2

Source: Javelin Strategy & Research; Council of Better Business Bureau, Inc.

Looking at the numbers presented in a table does not have the same impact as presenting numbers in a well-drawn chart or graph. The article did not include any graphs. This chapter will show you how to construct appropriate graphs to represent data and help you to get your point across to your audience.

See Statistics Today—Revisited at the end of the chapter for some suggestions on how to represent the data graphically.



© Image Source, all rights reserved. RF

OUTLINE

Introduction

2-1 Organizing Data

2-2 Histograms, Frequency Polygons, and Ogives

2-3 Other Types of Graphs

Summary

OBJECTIVES

After completing this chapter, you should be able to

- 1** Organize data using a frequency distribution.
- 2** Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.
- 3** Represent data using bar graphs, Pareto charts, time series graphs, pie graphs, and dotplots.
- 4** Draw and interpret a stem and leaf plot.

Introduction

When conducting a statistical study, the researcher must gather data for the particular variable under study. For example, if a researcher wishes to study the number of people who were bitten by poisonous snakes in a specific geographic area over the past several years, he or she has to gather the data from various doctors, hospitals, or health departments.

To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a *frequency distribution*.

After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study. The most useful method of presenting the data is by constructing *statistical charts* and *graphs*. There are many different types of charts and graphs, and each one has a specific purpose.

This chapter explains how to organize data by constructing frequency distributions and how to present the data by constructing charts and graphs. The charts and graphs illustrated here are histograms, frequency polygons, ogives, pie graphs, Pareto charts, and time series graphs. A graph that combines the characteristics of a frequency distribution and a histogram, called a stem and leaf plot, is also explained.

2-1 Organizing Data

OBJECTIVE 1

Organize data using a frequency distribution.

Suppose a researcher wished to do a study on the ages of the 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine*. When the data are in original form, they are called **raw data** and are listed next.

45	46	64	57	85
92	51	71	54	48
27	66	76	55	69
54	44	54	75	46
61	68	78	61	83
88	45	89	67	56
81	58	55	62	38
55	56	64	81	38
49	68	91	56	68
46	47	83	71	62

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a *frequency distribution*.

Unusual Stats

Of Americans 50 years old and over, 23% think their greatest achievements are still ahead of them.

A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.

Each raw data value is placed into a quantitative or qualitative category called a **class**. The **frequency** of a class then is the number of data values contained in a specific class. A frequency distribution is shown for the preceding data set.

Class limits	Tally	Frequency
27–35	/	1
36–44	///	3
45–53	/// ~~///~~	9
54–62	/// ~~///~~ ~~///~~	15
63–71	/// ~~///~~	10
72–80	///	3
81–89	/// //	7
90–98	//	2
		<u>50</u>

Now some general observations can be made from looking at the frequency distribution. For example, it can be stated that the majority of the wealthy people in the study are 45 years old or older.

The classes in this distribution are 27–35, 36–44, etc. These values are called *class limits*. The data values 27, 28, 29, 30, 31, 32, 33, 34, 35 can be tallied in the first class; 36, 37, 38, 39, 40, 41, 42, 43, 44 in the second class; and so on.

Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution*. The procedures for constructing these distributions are shown now.

Categorical Frequency Distributions

The **categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

EXAMPLE 2-1 Distribution of Blood Types

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

SOLUTION

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

The procedure for constructing a frequency distribution for categorical data is given next.

Step 1 Make a table as shown.

A Class	B Tally	C Frequency	D Percent
A			
B			
O			
AB			

Step 2 Tally the data and place the results in column B.

Step 3 Count the tallies and place the results in column C.

Step 4 Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100$$

where f = frequency of the class and n = total number of values. For example, in the class of type A blood, the percentage is

$$\% = \frac{5}{25} \cdot 100 = 20\%$$

Percentages are not normally part of a frequency distribution, but they can be added since they are used in certain types of graphs such as pie graphs. Also, the decimal equivalent of a percent is called a *relative frequency*.

Step 5 Find the totals for columns C (frequency) and D (percent). The completed table is shown. It is a good idea to add the percent column to make sure it sums to 100%. This column won't always sum to 100% because of rounding.

A Class	B Tally	C Frequency	D Percent
A		5	20
B		7	28
O		9	36
AB		4	16
		Total 25	100%

For the sample, more people have type O blood than any other type.

Grouped Frequency Distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a **grouped frequency distribution**. For example, a distribution of the blood glucose levels in milligrams per deciliter (mg/dL) for 50 randomly selected college students is shown.

Unusual Stats

Six percent of Americans say they find life dull.

Class limits	Class boundaries	Tally	Frequency
58–64	57.5–64.5	/	1
65–71	64.5–71.5	/	6
72–78	71.5–78.5		10
79–85	78.5–85.5		14
86–92	85.5–92.5		12
93–99	92.5–99.5		5
100–106	99.5–106.5		2
			Total 50

The procedure for constructing the preceding frequency distribution is given in Example 2–2; however, several things should be noted. In this distribution, the values 58 and 64 of the first class are called *class limits*. The **lower class limit** is 58; it represents the smallest data value that can be included in the class. The **upper class limit** is 64; it

represents the largest data value that can be included in the class. The numbers in the second column are called **class boundaries**. These numbers are used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 64 and 65.

Students sometimes have difficulty finding class boundaries when given the class limits. The basic rule of thumb is that *the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5*. For example, if the values in the data set are whole numbers, such as 59, 68, and 82, the limits for a class might be 58–64, and the boundaries are 57.5–64.5. Find the boundaries by subtracting 0.5 from 58 (the lower class limit) and adding 0.5 to 64 (the upper class limit).

$$\text{Lower limit} - 0.5 = 58 - 0.5 = 57.5 = \text{lower boundary}$$

$$\text{Upper limit} + 0.5 = 64 + 0.5 = 64.5 = \text{upper boundary}$$

Unusual Stats

One out of every hundred people in the United States is color-blind.

If the data are in tenths, such as 6.2, 7.8, and 12.6, the limits for a class hypothetically might be 7.8–8.8, and the boundaries for that class would be 7.75–8.85. Find these values by subtracting 0.05 from 7.8 and adding 0.05 to 8.8.

Class boundaries are not always included in frequency distributions; however, they give a more formal approach to the procedure of organizing data, including the fact that sometimes the data have been rounded. You should be familiar with boundaries since you may encounter them in a statistical study.

Finally, the **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example, the class width in the preceding distribution on the distribution of blood glucose levels is 7, found from $65 - 58 = 7$.

The class width can also be found by subtracting the lower boundary from the upper boundary for any given class. In this case, $64.5 - 57.5 = 7$.

Note: Do not subtract the limits of a single class. It will result in an incorrect answer.

The researcher must decide how many classes to use and the width of each class. To construct a frequency distribution, follow these rules:

1. *There should be between 5 and 20 classes.* Although there is no hard-and-fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data.
2. *It is preferable but not absolutely necessary that the class width be an odd number.* This ensures that the midpoint of each class has the same place value as the data. The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the first class in the example with glucose levels is

$$\frac{57.5 + 64.5}{2} = 61 \quad \text{or} \quad \frac{58 + 64}{2} = 61$$

The midpoint is the numeric location of the center of the class. Midpoints are necessary for graphing (see Section 2-2). If the class width is an even number, the

midpoint is in tenths. For example, if the class width is 6 and the boundaries are 5.5 and 11.5, the midpoint is

$$\frac{5.5 + 11.5}{2} = \frac{17}{2} = 8.5$$

Rule 2 is only a suggestion, and it is not rigorously followed, especially when a computer is used to group data.

3. *The classes must be mutually exclusive.* Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes. Many times, frequency distributions such as this

Age
10–20
20–30
30–40
40–50

are found in the literature or in surveys. If a person is 40 years old, into which class should she or he be placed? A better way to construct a frequency distribution is to use classes such as

Age
10–20
21–31
32–42
43–53

Recall that boundaries are mutually exclusive. For example, when a class boundary is 5.5 to 10.5, the data values that are included in that class are values from 6 to 10. A data value of 5 goes into the previous class, and a data value of 11 goes into the next-higher class.

4. *The classes must be continuous.* Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.
5. *The classes must be exhaustive.* There should be enough classes to accommodate all the data.
6. *The classes must be equal in width.* This avoids a distorted view of the data.

One exception occurs when a distribution has a class that is **open-ended**. That is, the first class has no specific lower limit, or the last class has no specific upper limit. A frequency distribution with an open-ended class is called an **open-ended distribution**. Here are two examples of distributions with open-ended classes.

Age	Frequency
10–20	3
21–31	6
32–42	4
43–53	10
54 and above	8

Minutes	Frequency
Below 110	16
110–114	24
115–119	38
120–124	14
125–129	5

The frequency distribution for age is open-ended for the last class, which means that anybody who is 54 years or older will be tallied in the last class. The distribution for minutes is open-ended for the first class, meaning that any minute values below 110 will be tallied in that class.

The steps for constructing a grouped frequency distribution are summarized in the following Procedure Table.

Procedure Table

Constructing a Grouped Frequency Distribution

- Step 1** Determine the classes.
- Find the highest and lowest values.
 - Find the range.
 - Select the number of classes desired.
 - Find the width by dividing the range by the number of classes and rounding up.
 - Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits.
 - Find the upper class limits.
 - Find the boundaries.
- Step 2** Tally the data.
- Step 3** Find the numerical frequencies from the tallies, and find the cumulative frequencies.

Example 2-2 shows the procedure for constructing a grouped frequency distribution, i.e., when the classes contain more than one data value.

EXAMPLE 2-2 Record High Temperatures

These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data, using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: *The World Almanac and Book of Facts*.

SOLUTION

The procedure for constructing a grouped frequency distribution for numerical data follows.

Step 1 Determine the classes.

Find the highest value and lowest value: $H = 134$ and $L = 100$.

Find the range: $R = \text{highest value} - \text{lowest value} = H - L$, so

$$R = 134 - 100 = 34$$

Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

Find the class width by dividing the range by the number of classes.

$$\text{Width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Unusual Stats

America's most popular beverages are soft drinks. It is estimated that, on average, each person drinks about 52 gallons of soft drinks per year, compared to 22 gallons of beer.

Historical Note

Florence Nightingale, a nurse in the Crimean War in 1854, used statistics to persuade government officials to improve hospital care of soldiers in order to reduce the death rate from unsanitary conditions in the military hospitals that cared for the wounded soldiers.

Round the answer up to the nearest whole number if there is a remainder: $4.9 \approx 5$. (Rounding *up* is different from rounding *off*. A number is rounded up if there is any decimal remainder when dividing. For example, $85 \div 6 = 14.167$ and is rounded up to 15. Also, $53 \div 4 = 13.25$ and is rounded up to 14. (Also, after dividing, if there is no remainder, you will need to add an extra class to accommodate all the data.)

Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. In this case, 100 is used. Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as shown, 100, 105, 110, etc.

Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

$$105 - 1 = 104$$

The first class is 100–104, the second class is 105–109, etc.

Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit:

$$99.5-104.5, 104.5-109.5, \text{ etc.}$$

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies.

The completed frequency distribution is

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	///	8
110–114	109.5–114.5	///	18
115–119	114.5–119.5	///	13
120–124	119.5–124.5	///	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			Total 50

The frequency distribution shows that the class 109.5–114.5 contains the largest number of temperatures (18) followed by the class 114.5–119.5 with 13 temperatures. Hence, most of the temperatures (31) fall between 110 and 119°F.

Sometimes it is necessary to use a *cumulative frequency distribution*. A **cumulative frequency distribution** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary). The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency. In this example, the cumulative frequency for the first class is $0 + 2 = 2$; for the second class it is $0 + 2 + 8 = 10$; for the third class it is $0 + 2 + 8 + 18 = 28$. Naturally, a shorter way to do this would be to just add the cumulative frequency of the class below to the frequency of the given class. For example, the cumulative frequency for the number of data values less than 114.5 can be

found by adding $10 + 18 = 28$. The cumulative frequency distribution for the data in this example is as follows:

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Cumulative frequencies are used to show how many data values are accumulated up to and including a specific class. In Example 2-2, of the total record high temperatures 28 are less than or equal to 114°F . Forty-eight of the total record high temperatures are less than or equal to 124°F .

After the raw data have been organized into a frequency distribution, it will be analyzed by looking for peaks and extreme values. The peaks show which class or classes have the most data values compared to the other classes. Extreme values, called *outliers*, show large or small data values that are relative to other data values.

When the range of the data values is relatively small, a frequency distribution can be constructed using single data values for each class. This type of distribution is called an **ungrouped frequency distribution** and is shown next.

EXAMPLE 2-3 Hours of Sleep

The data shown represent the number of hours 30 college students said they sleep per night. Construct and analyze a frequency distribution.

8	6	6	8	5	7
7	8	7	6	6	7
9	7	7	6	8	10
6	7	6	7	8	7
7	8	7	8	9	8

SOLUTION

Step 1 Determine the number of classes. Since the range is small ($10 - 5 = 5$), classes consisting of a single data value can be used. They are 5, 6, 7, 8, 9, and 10.

Note: If the data are continuous, class boundaries can be used. Subtract 0.5 from each class value to get the lower class boundary, and add 0.5 to each class value to get the upper class boundary.

Step 2 Tally the data.

Step 3 From the tallies, find the numerical frequencies and cumulative frequencies. The completed ungrouped frequency distribution is shown.

Class limits	Class boundaries	Tally	Frequency
5	4.5–5.5	/	1
6	5.5–6.5		7
7	6.5–7.5		11
8	7.5–8.5		8
9	8.5–9.5		2
10	9.5–10.5	/	1

In this case, 11 students sleep 7 hours a night. Most of the students sleep between 5.5 and 8.5 hours.

The cumulative frequencies are

	Cumulative frequency
Less than 4.5	0
Less than 5.5	1
Less than 6.5	8
Less than 7.5	19
Less than 8.5	27
Less than 9.5	29
Less than 10.5	30

Interesting Fact

Male dogs bite children more often than female dogs do; however, female cats bite children more often than male cats do.

When you are constructing a frequency distribution, the guidelines presented in this section should be followed. However, you can construct several different but correct frequency distributions for the same data by using a different class width, a different number of classes, or a different starting point.

Furthermore, the method shown here for constructing a frequency distribution is not unique, and there are other ways of constructing one. Slight variations exist, especially in computer packages. But regardless of what methods are used, classes should be mutually exclusive, continuous, exhaustive, and of equal width.

In summary, the different types of frequency distributions were shown in this section. The first type, shown in Example 2–1, is used when the data are categorical (nominal), such as blood type or political affiliation. This type is called a categorical frequency distribution. The second type of distribution is used when the range is large and classes several units in width are needed. This type is called a grouped frequency distribution and is shown in Example 2–2. Another type of distribution is used for numerical data and when the range of data is small, as shown in Example 2–3. Since each class is only one unit, this distribution is called an ungrouped frequency distribution.

All the different types of distributions are used in statistics and are helpful when one is organizing and presenting data.

The reasons for constructing a frequency distribution are as follows:

1. To organize the data in a meaningful, intelligible way.
2. To enable the reader to determine the nature or shape of the distribution.
3. To facilitate computational procedures for measures of average and spread (shown in Sections 3–1 and 3–2).

4. To enable the researcher to draw charts and graphs for the presentation of data (shown in Section 2-2).
5. To enable the reader to make comparisons among different data sets.

The factors used to analyze a frequency distribution are essentially the same as those used to analyze histograms and frequency polygons, which are shown in Section 2-2.

Applying the Concepts 2-1

Ages of Presidents at Inauguration

The data represent the ages of our Presidents at the time they were first inaugurated.

57	61	57	57	58	57	61	54	68
51	49	64	50	48	65	52	56	46
54	49	51	47	55	55	54	42	51
56	55	51	54	51	60	62	43	55
56	61	52	69	64	46	54	47	

1. Were the data obtained from a population or a sample? Explain your answer.
2. What was the age of the oldest President?
3. What was the age of the youngest President?
4. Construct a frequency distribution for the data. (Use your own judgment as to the number of classes and class size.)
5. Are there any peaks in the distribution?
6. Identify any possible outliers.
7. Write a brief summary of the nature of the data as shown in the frequency distribution.

See page 108 for the answers.

Exercises 2-1

1. List five reasons for organizing data into a frequency distribution.
2. Name the three types of frequency distributions, and explain when each should be used.
3. How many classes should frequency distributions have? Why should the class width be an odd number?
4. What are open-ended frequency distributions? Why are they necessary?

For Exercises 5–8, find the class boundaries, midpoints, and widths for each class.

5. 58–62
6. 125–131
7. 16.35–18.46
8. 16.3–18.5

For Exercises 9–12, show frequency distributions that are incorrectly constructed. State the reasons why they are wrong.

9. Class	Frequency
10–19	1
20–29	2
30–34	0
35–45	5
46–51	8

10. Class	Frequency
5–9	1
9–13	2
13–17	5
17–20	6
20–24	3

11. Class Frequency

162–164	3
165–167	7
168–170	18
174–176	0
177–179	5

12. Class Frequency

9–13	1
14–19	6
20–25	2
26–28	5
29–32	9

- 13. Favorite Coffee Flavor** A survey was taken asking the favorite flavor of a coffee drink a person prefers. The responses were V = Vanilla, C = Caramel, M = Mocha, H = Hazelnut, and P = Plain. Construct a categorical frequency distribution for the data. Which class has the most data values and which class has the fewest data values?

V C P P M M P P M C
M M V M M M V M M M
P V C M V M C P M P
M M M P M M C V M C
C P M P M H H P H P

- 14. Trust in Internet Information** A survey was taken on how much trust people place in the information they read on the Internet. Construct a categorical frequency distribution for the data. A = trust in all that they read, M = trust in most of what they read, H = trust in about one-half of what they read, S = trust in a small portion of what they read. (Based on information from the *UCLA Internet Report*.)

M M M A H M S M H M
S M M M M A M M A M
M M H M M M H M H M
A M M M H M M M M M

- 15. Eating at Fast Food Restaurants** A survey was taken of 50 individuals. They were asked how many days per week they ate at a fast-food restaurant. Construct a frequency distribution using 8 classes (0–7). Based on the distribution, how often did most people eat at a fast-food restaurant?

1 3 4 0 4
5 2 2 3 1
2 2 2 2 2
2 2 2 2 3
2 2 5 2 4
2 4 5 2 1
4 1 3 2 2
2 0 7 2 3
2 2 2 5 2
3 3 4 1 3

- 16. Ages of Dogs** The ages of 20 dogs in a pet shelter are shown. Construct a frequency distribution using 7 classes.

5 8 7 6 3
9 4 4 5 8
7 4 7 5 7
3 5 8 4 9

- 17. Maximum Wind Speeds** The data show the maximum wind speeds in miles per hour recorded for 40 states. Construct a frequency distribution using 7 classes.

59 78 62 72 67
76 92 77 64 83
64 70 67 75 75
78 75 71 72 93
68 69 76 72 85
64 70 77 74 72
53 67 48 76 59
87 53 77 70 63

Source: NOAA

- 18. Stories in the World's Tallest Buildings** The number of stories in each of a sample of the world's 30 tallest buildings follows. Construct a grouped frequency distribution and a cumulative frequency distribution with 7 classes.

88 88 110 88 80 69 102 78 70 55
79 85 80 100 60 90 77 55 75 55
54 60 75 64 105 56 71 70 65 72

Source: *New York Times Almanac*.

- 19. Ages of Declaration of Independence Signers** The ages of the signers of the Declaration of Independence are shown. (Age is approximate since only the birth year appeared in the source, and one has been omitted since his birth year is unknown.) Construct a grouped frequency distribution and a cumulative frequency distribution for the data, using 7 classes.

41 54 47 40 39 35 50 37 49 42 70 32
44 52 39 50 40 30 34 69 39 45 33 42
44 63 60 27 42 34 50 42 52 38 36 45
35 43 48 46 31 27 55 63 46 33 60 62
35 46 45 34 53 50 50

Source: *The Universal Almanac*.

- 20. Salaries of Governors** Here are the salaries (in dollars) of the governors of 25 randomly selected states. Construct a grouped frequency distribution with 6 classes.

112,895 117,312 140,533 110,000 115,331
95,000 177,500 120,303 139,590 150,000
173,987 130,000 133,821 144,269 142,542
150,000 145,885 105,000 93,600 166,891
130,273 70,000 113,834 117,817 137,092

Source: *World Almanac*.

- 21. Charity Donations** A random sample of 30 large companies in the United States shows the amount,

in millions of dollars, that each company donated to charity for a specific year. Construct a frequency distribution for the data, using 9 classes.

26	25	19	31	14
48	35	43	25	46
17	21	57	58	34
41	12	27	15	53
16	63	82	23	52
56	75	19	26	88

- 22. Unclaimed Expired Prizes** The number of unclaimed expired prizes (in millions of dollars) for lottery tickets bought in a sample of states is shown. Construct a frequency distribution for the data, using 5 classes.

28.5	51.7	19	5
2	1.2	14	14.6
0.8	11.6	3.5	30.1
1.7	1.3	13	14

- 23. Scores in the Rose Bowl** The data show the scores of the winning teams in the Rose Bowl. Construct a frequency distribution for the data using a class width of 7.

24	20	45	21	26	38	49	32	41	38
28	34	37	34	17	38	21	20	41	38
21	38	34	46	17	22	20	22	45	20
45	24	28	23	17	17	27	14	23	18

Source: The World Almanac.

- 24. Consumption of Natural Gas** Construct a frequency distribution for the energy consumption of natural gas (in billions of Btu) by the 50 states and the District of Columbia. Use 9 classes.

474	475	205	639	197	344	3	409	247	66
377	87	747	1166	223	248	958	406	251	3462
2391	514	371	58	224	530	317	267	769	9
188	289	76	678	331	52	214	165	255	319
34	1300	284	834	114	1082	73	62	95	393
146									

Source: Time Almanac.

- 25. Average Wind Speeds** A sample of 40 large cities was selected, and the average of the wind speeds was computed for each city over one year. Construct a frequency distribution, using 7 classes.

12.2	9.1	11.2	9.0
10.5	8.2	8.9	12.2
9.5	10.2	7.1	11.0
6.2	7.9	8.7	8.4
8.9	8.8	7.1	10.1
8.7	10.5	10.2	10.7
7.9	8.3	8.7	8.7
10.4	7.7	12.3	10.7
7.7	7.8	11.8	10.5
9.6	9.6	8.6	10.3

Source: World Almanac and Book of Facts.

- 26. Percentage of People Who Completed 4 or More Years of College** Listed by state are the percentages of the population who have completed 4 or more years of a college education. Construct a frequency distribution with 7 classes.

21.4	26.0	25.3	19.3	29.5	35.0	34.7	26.1	25.8	23.4
27.1	29.2	24.5	29.5	22.1	24.3	28.8	20.0	20.4	26.7
35.2	37.9	24.7	31.0	18.9	24.5	27.0	27.5	21.8	32.5
33.9	24.8	31.7	25.6	25.7	24.1	22.8	28.3	25.8	29.8
23.5	25.0	21.8	25.2	28.7	33.6	33.6	30.3	17.3	25.4

Source: New York Times Almanac.

Extending the Concepts

- 27. JFK Assassination** A researcher conducted a survey asking people if they believed more than one person was involved in the assassination of John F. Kennedy. The results were as follows: 73% said yes, 19% said no, and 9% had no opinion. Is there anything suspicious about the results?

- 28. The Value of Pi** The ratio of the circumference of a circle to its diameter is known as π (pi). The value of π is an irrational number, which means that the decimal

part goes on forever and there is no fixed sequence of numbers that repeats. People have found the decimal part of π to over a million places. We can statistically study the number. Shown here is the value of π to 40 decimal places. Construct an ungrouped frequency distribution for the digits. Based on the distribution, do you think each digit appears equally in the number?

3.1415926535897932384626433832795028841971

Technology

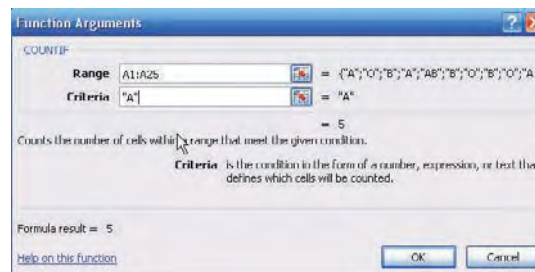
EXCEL

Step by Step

Step by Step

Categorical Frequency Table (Qualitative or Discrete Data)

1. In an open workbook, select cell A1 and type in all the blood types from Example 2–1 down column A.
2. Type in the variable name **Blood Type** in cell B1.
3. Select cell B2 and type in the four different blood types down the column.
4. Type in the name **Count** in cell C1.
5. Select cell C2. From the toolbar, select the Formulas tab on the toolbar.
6. Select the Insert Function icon , then select the Statistical category in the Insert Function dialog box.
7. Select the Countif function from the function name list.
8. In the dialog box, type **A1:A25** in the **Range** box. Type in the blood type “A” in quotes in the **Criteria** box. The count or frequency of the number of data corresponding to the blood type should appear below the input. Repeat for the remaining blood types.
9. After all the data have been counted, select cell C6 in the worksheet.
10. From the toolbar select Formulas, then AutoSum and type in C2:C5 to insert the total frequency into cell C6.



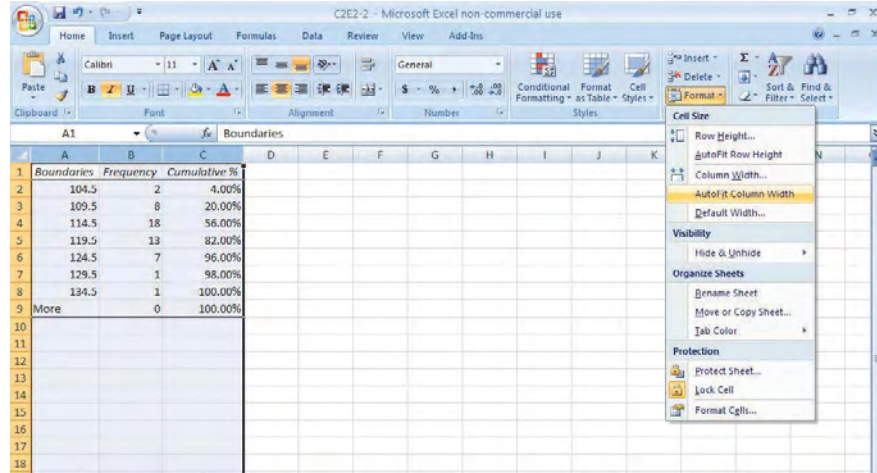
After entering data or a heading into a worksheet, you can change the width of a column to fit the input. To automatically change the width of a column to fit the data:

1. Select the column or columns that you want to change.
2. On the Home tab, in the Cells group, select Format.
3. Under Cell Size, click Autofit Column Width.

Making a Grouped Frequency Distribution (Quantitative Data)

1. Press [Ctrl]-N for a new workbook.
2. Enter the raw data from Example 2–2 in column A, one number per cell.
3. Enter the upper class boundaries in column B.
4. From the toolbar select the Data tab, then click Data Analysis.
5. In the Analysis Tools, select Histogram and click [OK].
6. In the Histogram dialog box, type **A1:A50** in the Input Range box and type **B1:B7** in the Bin Range box.
7. Select New Worksheet Ply, and check the Cumulative Percentage option. Click [OK].
8. You can change the label for the column containing the upper class boundaries and expand the width of the columns automatically after relabeling:
Select the Home tab from the toolbar.

Highlight the columns that you want to change.
Select Format, then AutoFit Column Width.



Note: By leaving the Chart Output unchecked, a new worksheet will display the table only.

MINITAB Step by Step

Make a Categorical Frequency Table (Qualitative or Discrete Data)

1. Type in all the blood types from Example 2-1 down C1 of the worksheet.
A B B AB O O O B AB B B B O A O A O O O AB AB A O B A
2. Click above row 1 and name the column **BloodType**.
3. Select **Stat>Tables>Tally Individual Values**.
The cursor should be blinking in the Variables dialog box. If not, click inside the dialog box.
4. Double-click C1 in the Variables list.
5. Check the boxes for the statistics: Counts, Percents, and Cumulative percents.
6. Click [OK]. The results will be displayed in the Session Window as shown.

Tally for Discrete Variables: BloodType

BloodType	Count	Percent	CumPct
A	5	20.00	20.00
AB	4	16.00	36.00
B	7	28.00	64.00
O	9	36.00	100.00
N=	25		

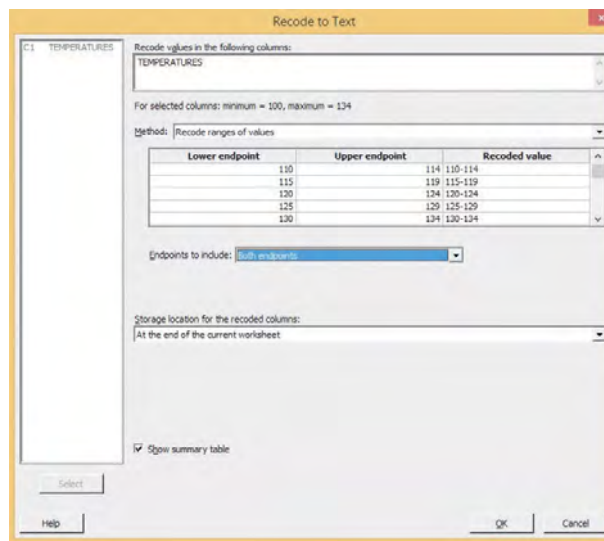
Make a Grouped Frequency Distribution (Quantitative Variable)

1. Select **File>New>Minitab Worksheet**. A new worksheet will be added to the project.
2. Type the data used in Example 2-2 into C1. Name the column **TEMPERATURES**.
3. Use the instructions in the textbook to determine the class limits of 100 to 134 in increments of 5.
In the next step you will create a new column of data, converting the numeric variable to text categories that can be tallied.

4. Select **Data>Recode>to Text**.

- The cursor should be blinking in Recode values in the following columns. If not, click inside the box, then double-click C1 Temperatures in the list. Only quantitative variables will be shown in this list.
- Click inside the Method: box and select **Recode ranges of values**.
- Press [Tab] to move to the table.
- Type 100 in the Lower endpoint column, press [Tab], type 104 in the Upper endpoint column.
- Press [Tab] to move to the Recoded value column, and type the text category **100–104**.
- Continue to tab to each dialog box, typing the lower endpoint and upper endpoint and then the category until the last category has been entered.
- Click inside the Endpoints to include: box and select **Both endpoints**.

The dialog box should look like the one shown.



- Click [OK]. In the worksheet, a new column of data will be created in the first empty column, C2. This new variable will contain the category for each value in C1. The column C2-T contains alphanumeric data.
- Click **Stat>Tables>Tally Individual Values**, then double-click Recoded TEMPERATURES in the Variables list.
 - Check the boxes for the desired statistics, such as Counts, Percents, and Cumulative percents.
 - Click [OK].

The table will be displayed in the Session Window. Eighteen states have high temperatures between 110 and 114°F. Eighty-two percent of the states have record high temperatures less than or equal to 119°F.

Tally for Discrete Variables: Recoded TEMPERATURES

Recoded TEMPERATURES	Count	Percent	CumPct
100–104	2	4.00	4.00
105–109	8	16.00	20.00
110–114	18	36.00	56.00
115–119	13	26.00	82.00
120–124	7	14.00	96.00
125–129	1	2.00	98.00
130–134	1	2.00	100.00
N=	50		

- Click **File>Save Project As . . .**, and type the name of the project file, **Ch2-1**. This will save the two worksheets and the Session Window.

2–2 Histograms, Frequency Polygons, and Ogives

OBJECTIVE 2

Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The three most commonly used graphs in research are

1. The histogram.
2. The frequency polygon.
3. The cumulative frequency graph, or ogive (pronounced o-jive).

The steps for constructing the histogram, frequency polygon, and the ogive are summarized in the procedure table.

Procedure Table

Constructing a Histogram, Frequency Polygon, and Ogive

Step 1 Draw and label the x and y axes.

Step 2 On the x axis, label the class boundaries of the frequency distribution for the histogram and ogive. Label the midpoints for the frequency polygon.

Step 3 Plot the frequencies for each class, and draw the vertical bars for the histogram and the lines for the frequency polygon and ogive.

(Note: Remember that the lines for the frequency polygon begin and end on the x axis while the lines for the ogive begin on the x axis.)

Historical Note

Karl Pearson introduced the histogram in 1891. He used it to show time concepts of various reigns of Prime Ministers.

The Histogram

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

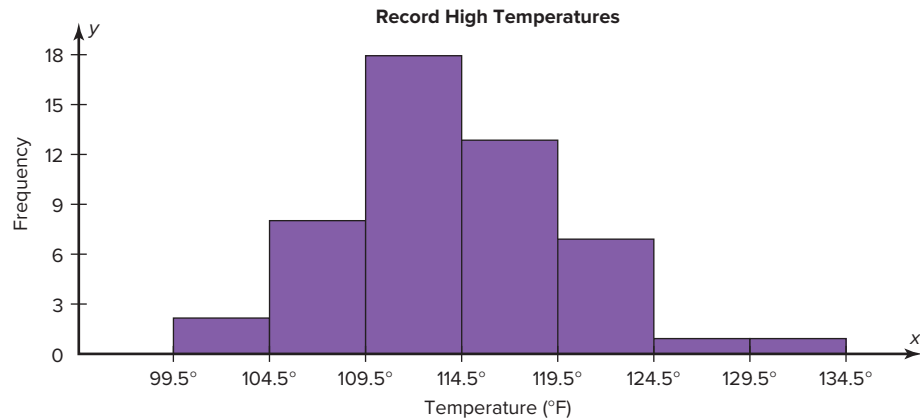
EXAMPLE 2–4 Record High Temperatures

Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states (see Example 2–2).

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

SOLUTION

- Step 1** Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.
- Step 2** Represent the frequency on the y axis and the class boundaries on the x axis.
- Step 3** Using the frequencies as the heights, draw vertical bars for each class. See Figure 2–1.

FIGURE 2–1 Histogram for Example 2–4

As the histogram shows, the class with the greatest number of data values (18) is 109.5–114.5, followed by 13 for 114.5–119.5. The graph also has one peak with the data clustering around it.

Historical Note

Graphs originated when ancient astronomers drew the position of the stars in the heavens. Roman surveyors also used coordinates to locate landmarks on their maps.

The development of statistical graphs can be traced to William Playfair (1759–1823), an engineer and drafter who used graphs to present economic data pictorially.

The Frequency Polygon

Another way to represent the same data set is by using a frequency polygon.

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Example 2–5 shows the procedure for constructing a frequency polygon. Be sure to begin and end on the x axis.

EXAMPLE 2–5 Record High Temperatures

Using the frequency distribution given in Example 2–4, construct a frequency polygon.

SOLUTION

- Step 1** Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

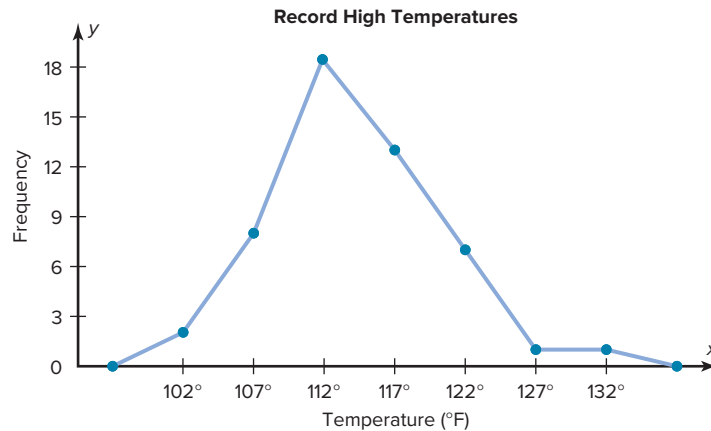
$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

- Step 2** Draw the x and y axes. Label the x axis with the midpoint of each class, and then use a suitable scale on the y axis for the frequencies.
- Step 3** Using the midpoints for the x values and the frequencies as the y values, plot the points.
- Step 4** Connect adjacent points with line segments. Draw a line back to the x axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located, as shown in Figure 2-2.

FIGURE 2-2
Frequency Polygon for
Example 2-5



The frequency polygon and the histogram are two different ways to represent the same data set. The choice of which one to use is left to the discretion of the researcher.

The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the *cumulative frequency graph*, or *ogive*. The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

Example 2-6 shows the procedure for constructing an ogive. Be sure to start on the x axis.

EXAMPLE 2-6 Record High Temperatures

Construct an ogive for the frequency distribution described in Example 2-4.

SOLUTION

Step 1 Find the cumulative frequency for each class.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Step 2 Draw the x and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies. (Depending on the numbers in the cumulative frequency columns, scales such as 0, 1, 2, 3, . . . , or 5, 10, 15, 20, . . . , or 1000, 2000, 3000, . . . can be used. Do *not* label the y axis with the numbers in the cumulative frequency column.) In this example, a scale of 0, 5, 10, 15, . . . will be used.

Step 3 Plot the cumulative frequency at each upper class boundary, as shown in Figure 2–3. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

Step 4 Starting with the first upper class boundary, 104.5, connect adjacent points with line segments, as shown in Figure 2–4. Then extend the graph to the first lower class boundary, 99.5, on the x axis.

FIGURE 2–3

Plotting the Cumulative Frequency for Example 2–6

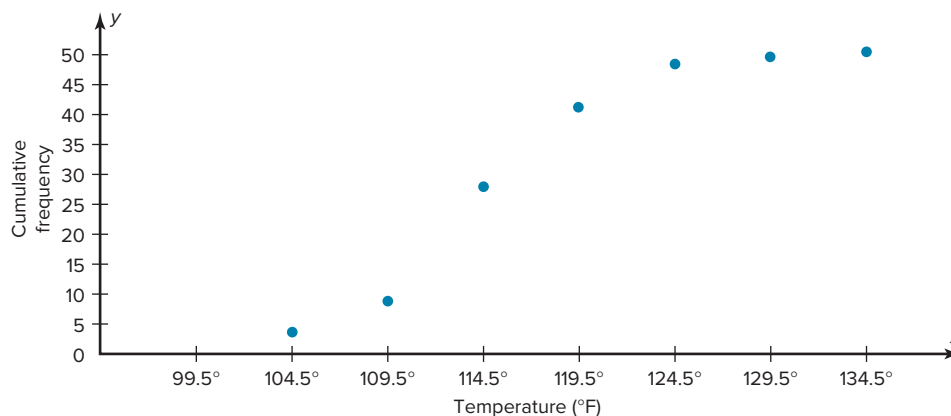


FIGURE 2–4

Ogive for Example 2–6

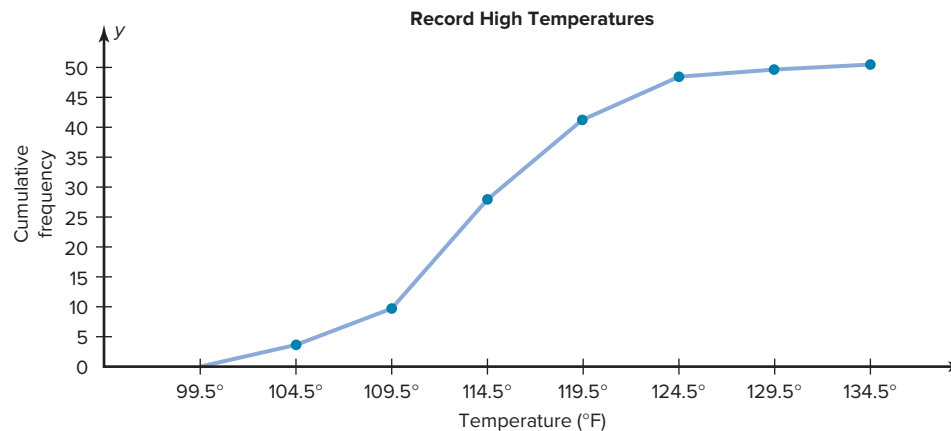
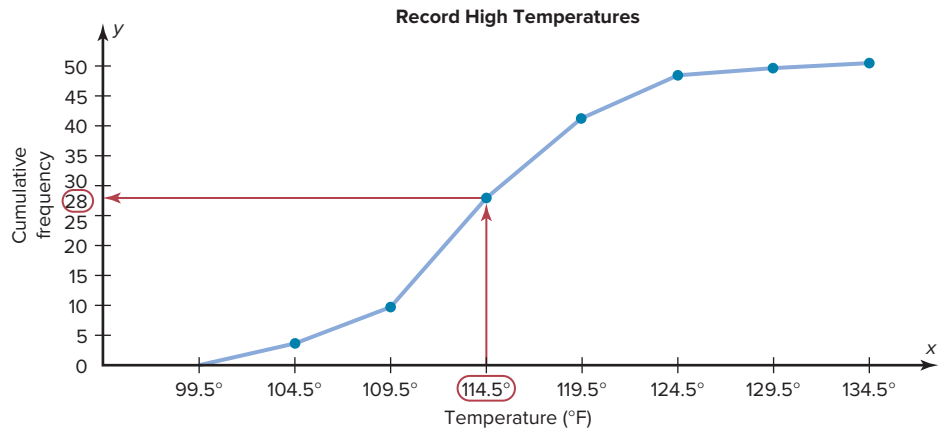


FIGURE 2-5

Finding a Specific Cumulative Frequency

**Unusual Stats**

Twenty-two percent of Americans sleep 6 hours a day or less.

Cumulative frequency graphs are used to visually represent how many values are below a certain upper class boundary. For example, to find out how many record high temperatures are less than 114.5°F, locate 114.5°F on the x axis, draw a vertical line up until it intersects the graph, and then draw a horizontal line at that point to the y axis. The y axis value is 28, as shown in Figure 2-5.

Relative Frequency Graphs

The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs**.

Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class. For example, if you wanted to compare the age distribution of adults in Philadelphia, Pennsylvania, with the age distribution of adults of Erie, Pennsylvania, you would use relative frequency distributions. The reason is that since the population of Philadelphia is 1,526,006 and the population of Erie is 101,786, the bars using the actual data values for Philadelphia would be much taller than those for the same classes for Erie.

To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1. These graphs are similar to the ones that use raw data as frequencies, but the values on the y axis are in terms of proportions. Example 2-7 shows the three types of relative frequency graphs.

EXAMPLE 2-7 Ages of State Governors

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution shown. This is a grouped frequency distribution using the ages (at the time of this writing) of the governors of the 50 states of the United States.

Class boundaries	Frequency
42.5–47.5	4
47.5–52.5	4
52.5–57.5	11
57.5–62.5	14
62.5–67.5	9
67.5–72.5	5
72.5–77.5	3
	Total 50

SOLUTION

Step 1 Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

For the class 42.5–47.5 the relative frequency = $\frac{4}{50} = 0.08$; for the class 47.5–52.5, the relative frequency is $\frac{4}{50} = 0.08$; for the class 52.5–57.5, the relative frequency is $\frac{11}{50} = 0.22$, and so on.

Place these values in the column labeled Relative Frequency. Also, find the midpoints, as shown in Example 2-5, for each class and place them in the midpoint column

Class boundaries	Midpoints	Relative frequency
42.5–47.5	45	0.08
47.5–52.5	50	0.08
52.5–57.5	55	0.22
57.5–62.5	60	0.28
62.5–67.5	65	0.18
67.5–72.5	70	0.10
72.5–77.5	75	0.06

Step 2 Find the cumulative relative frequencies. To do this, add the frequency in each class to the total frequency of the preceding class. In this case, $0.00 + 0.08 = 0.08$, $0.08 + 0.08 = 0.16$, $0.16 + 0.22 = 0.38$, $0.38 + 0.28 = 0.66$, etc. Place these values in a column labeled Cumulative relative frequency.

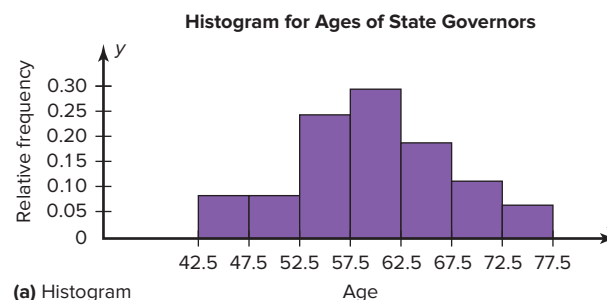
An alternative method would be to change the cumulative frequencies for the classes to relative frequencies. (Divide each by the total).

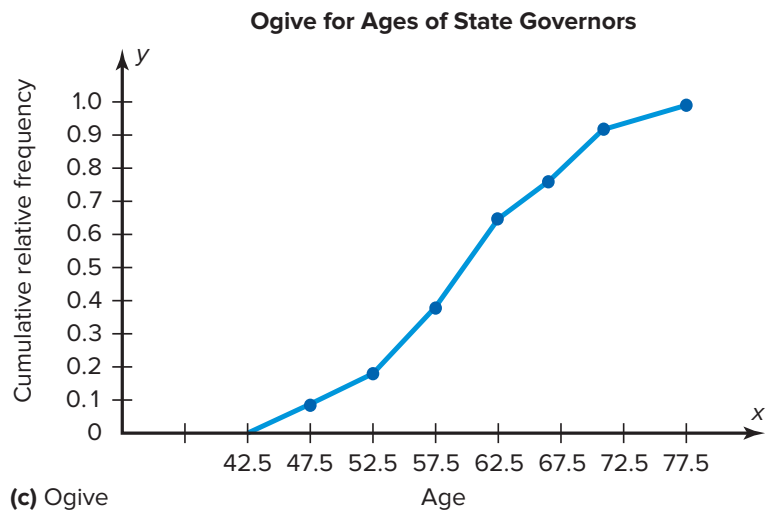
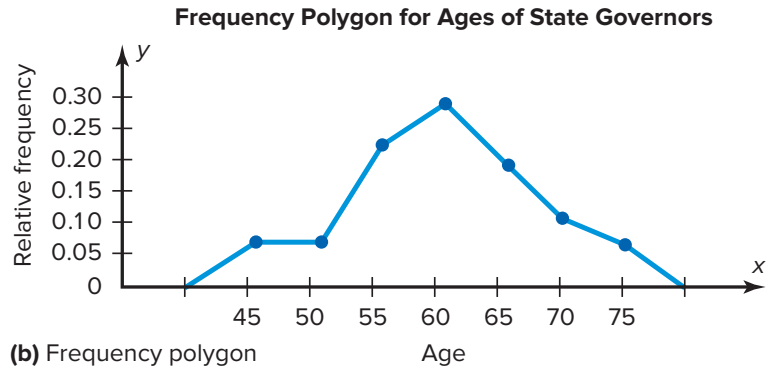
	Cumulative frequency	Cumulative relative frequency
Less than 42.5	0	0.00
Less than 47.5	4	0.08
Less than 52.5	8	0.16
Less than 57.5	19	0.38
Less than 62.5	33	0.66
Less than 67.5	42	0.84
Less than 72.5	47	0.94
Less than 77.5	50	1.00

Step 3 Draw each graph as shown in Figure 2-6. For the histogram and ogive, use the class boundaries along the x axis. For the frequency, use the midpoints on the x axis. For the scale on the y axis, use proportions.

FIGURE 2-6

Graphs for Example 2-7





Distribution Shapes

When one is describing data, it is important to be able to recognize the shapes of the distribution values. In later chapters, you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.

A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution. Several of the most common shapes are shown in Figure 2-7: *the bell-shaped or mound-shaped, the uniform-shaped, the J-shaped, the reverse J-shaped, the positively or right-skewed shape, the negatively or left-skewed shape, the bimodal-shaped, and the U-shaped.*

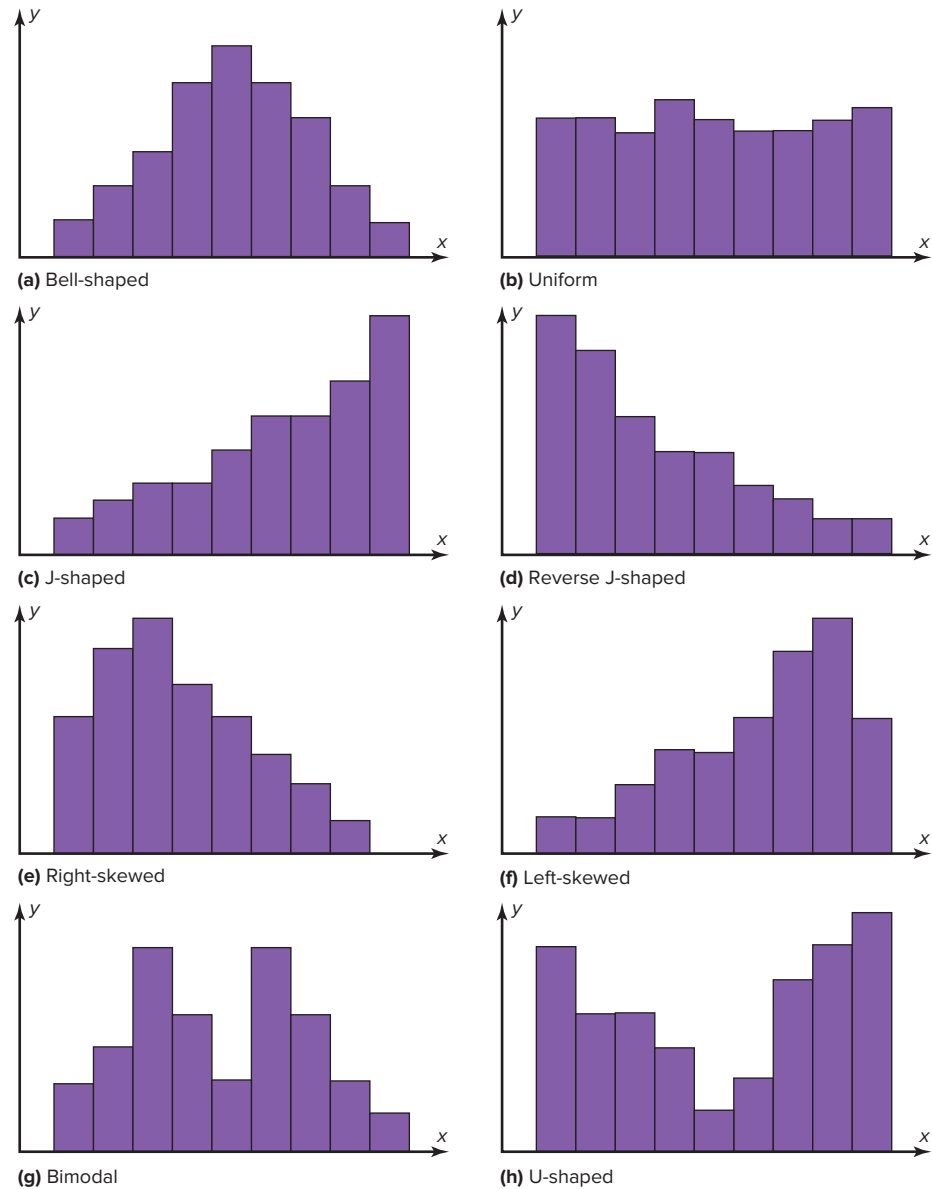
Distributions are most often not perfectly shaped, so it is not necessary to have an exact shape but rather to identify an overall pattern.

A *bell-shaped distribution* shown in Figure 2-7(a) has a single peak and tapers off at either end. It is approximately symmetric; i.e., it is roughly the same on both sides of a line running through the center.

A *uniform distribution* is basically flat or rectangular. See Figure 2-7(b).

A *J-shaped distribution* is shown in Figure 2-7(c), and it has a few data values on the left side and increases as one moves to the right. A *reverse J-shaped distribution* is the opposite of the J-shaped distribution. See Figure 2-7(d).

When the peak of a distribution is to the left and the data values taper off to the right, a distribution is said to be *positively or right-skewed*. See Figure 2-7(e). When

FIGURE 2-7
Distribution Shapes

the data values are clustered to the right and taper off to the left, a distribution is said to be *negatively or left-skewed*. See Figure 2-7(f). Skewness will be explained in detail in Chapter 3. Distributions with one peak, such as those shown in Figure 2-7(a), (e), and (f), are said to be *unimodal*. (The highest peak of a distribution indicates where the mode of the data values is. The mode is the data value that occurs more often than any other data value. Modes are explained in Chapter 3.) When a distribution has two peaks of the same height, it is said to be *bimodal*. See Figure 2-7(g). Finally, the graph shown in Figure 2-7(h) is a *U-shaped* distribution.

Distributions can have other shapes in addition to the ones shown here; however, these are some of the more common ones that you will encounter in analyzing data.

When you are analyzing histograms and frequency polygons, look at the shape of the curve. For example, does it have one peak or two peaks? Is it relatively flat, or is it U-shaped? Are the data values spread out on the graph, or are they clustered around the center? Are there data values in the extreme ends? These may be *outliers*. (See Section 3-3 for an explanation of outliers.) Are there any gaps in the histogram, or does

the frequency polygon touch the x axis somewhere other than at the ends? Finally, are the data clustered at one end or the other, indicating a *skewed distribution*?

For example, the histogram for the record high temperatures in Figure 2–1 shows a single peaked distribution, with the class 109.5–114.5 containing the largest number of temperatures. The distribution has no gaps, and there are fewer temperatures in the highest class than in the lowest class.

Applying the Concepts 2–2

Selling Real Estate

Assume you are a realtor in Bradenton, Florida. You have recently obtained a listing of the selling prices of the homes that have sold in that area in the last 6 months. You wish to organize those data so you will be able to provide potential buyers with useful information. Use the following data to create a histogram, frequency polygon, and cumulative frequency polygon.

142,000	127,000	99,600	162,000	89,000	93,000	99,500
73,800	135,000	119,500	67,900	156,300	104,500	108,650
123,000	91,000	205,000	110,000	156,300	104,000	133,900
179,000	112,000	147,000	321,550	87,900	88,400	180,000
159,400	205,300	144,400	163,000	96,000	81,000	131,000
114,000	119,600	93,000	123,000	187,000	96,000	80,000
231,000	189,500	177,600	83,400	77,000	132,300	166,000

1. What questions could be answered more easily by looking at the histogram rather than the listing of home prices?
2. What different questions could be answered more easily by looking at the frequency polygon rather than the listing of home prices?
3. What different questions could be answered more easily by looking at the cumulative frequency polygon rather than the listing of home prices?
4. Are there any extremely large or extremely small data values compared to the other data values?
5. Which graph displays these extremes the best?
6. Is the distribution skewed?

See page 108 for the answers.

Exercises 2–2

- 1. Do Students Need Summer Development?** For 108 randomly selected college applicants, the following frequency distribution for entrance exam scores was obtained. Construct a histogram, frequency polygon, and ogive for the data. (The data for this exercise will be used for Exercise 13 in this section.)

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9
Total	108

Applicants who score above 107 need not enroll in a summer developmental program. In this group, how many students do not have to enroll in the developmental program?

- 2. Bear Kills** The number of bears killed in 2014 for 56 counties in Pennsylvania is shown in the frequency distribution. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. How many counties had 75 or fewer bears killed? (The data for this exercise will be used for Exercise 14 of this section.)

Class limits	Frequency
1–25	16
26–50	14
51–75	9
76–100	8
101–125	5
126–150	0
151–175	1
176–200	1
201–225	0
226–250	0
251–275	2
Total	56

Source: Pennsylvania State Game Commission.

- 3. Pupils Per Teacher** The average number of pupils per teacher in each state is shown. Construct a grouped frequency distribution with 6 classes. Draw a histogram, frequency polygon, and ogive. Analyze the distribution.

16	16	15	12	14
13	16	14	15	14
18	18	18	12	15
15	16	16	15	15
25	19	15	12	22
18	14	13	17	9
13	14	13	16	12
14	16	10	22	20
12	14	18	15	14
16	12	12	13	15

Source: U.S. Department of Education.

- 4. Number of College Faculty** The number of faculty listed for a sample of private colleges that offer only bachelor's degrees is listed below. Use these data to construct a frequency distribution with 7 classes, a histogram, a frequency polygon, and an ogive. Discuss the shape of this distribution. What proportion of schools have 180 or more faculty?

165	221	218	206	138	135	224	204
70	210	207	154	155	82	120	116
176	162	225	214	93	389	77	135
221	161	128	310				

Source: World Almanac and Book of Facts.

- 5. Railroad Crossing Accidents** The data show the number of railroad crossing accidents for the 50 states of the United States for a specific year. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. (The data in this exercise will be used for Exercise 15 in this section.)

Class limits	Frequency
1–43	24
44–86	17
87–129	3
130–172	4
173–215	1
216–258	0
259–301	0
302–344	1
Total	50

Source: Federal Railroad Administration.

- 6. NFL Salaries** The salaries (in millions of dollars) for 31 NFL teams for a specific season are given in this frequency distribution.

Construct a histogram, a frequency polygon, and an ogive for the data; and comment on the shape of the distribution. (The data for this exercise will be used for Exercise 16 of this section.)

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	5
Total	31

Source: NFL.com

- 7. Suspension Bridges Spans** The following frequency distribution shows the length (in feet) of the main spans of the longest suspension bridges in the United States. Construct a histogram, frequency polygon, and ogive for the distribution. Describe the shape of the distribution.

Class limits	Frequency
1260–1734	12
1735–2209	6
2210–2684	3
2685–3159	1
3160–3634	1
3635–4109	1
4110–4584	2

Source: U.S. Department of Transportation.

- 8. Costs of Utilities** The frequency distribution represents the cost (in cents) for the utilities of states that supply much of their own power. Construct a histogram, frequency polygon, and ogive for the data. Is the distribution skewed?

Class limits	Frequency
6–8	12
9–11	16
12–14	3
15–17	1
18–20	0
21–23	0
24–26	1
Total	33

- 9. Air Pollution** One of the air pollutants that is measured in selected cities is sulfur dioxide. This pollutant occurs when fossil fuels are burned. This pollutant is measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$). The results obtained from a sample of 24 cities are shown in the frequency distributions. One sample was taken recently, and the other sample of the same cities was taken

5 years ago. Construct a histogram and compare the two distributions.

Class limits	Frequency (now)	Frequency (5 years ago)
10–14	6	5
15–19	4	4
20–24	3	2
25–29	2	3
30–34	5	6
35–39	1	2
40–44	2	1
45–49	1	1
	Total 24	Total 24

- 10. Making the Grade** The frequency distributions shown indicate the percentages of public school students in fourth-grade reading and mathematics who performed at or above the required proficiency levels for the 50 states in the United States. Draw histograms for each, and decide if there is any difference in the performance of the students in the subjects.

Class	Reading frequency	Math frequency
17.5–22.5	7	5
22.5–27.5	6	9
27.5–32.5	14	11
32.5–37.5	19	16
37.5–42.5	3	8
42.5–47.5	1	1
	Total 50	Total 50

Source: National Center for Educational Statistics.

- 11. Blood Glucose Levels** The frequency distribution shows the blood glucose levels (in milligrams per deciliter) for 50 patients at a medical facility. Construct a histogram, frequency polygon, and ogive for the data. Comment on the shape of the distribution. What range of glucose levels did most patients fall into?

Class limits	Frequency
60–64	2
65–69	1
70–74	5
75–79	12
80–84	18
85–89	6
90–94	5
95–99	1
	Total 50

- 12. Waiting Times** The frequency distribution shows the waiting times (in minutes) for 50 patients at a walk-in medical facility. Construct a histogram, frequency polygon, and ogive for the data. Is the

distribution skewed? How many patients waited longer than 30 minutes?

Class limits	Frequency
11–15	7
16–20	9
21–25	15
26–30	9
31–35	5
36–40	3
41–45	2
	Total 50

- 13.** Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 1 of this section.
- 14.** Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 2 of this section.
- 15.** Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 5 of this section.
- 16.** Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 6 of this section.
- 17. Home Runs** The data show the most number of home runs hit by a batter in the American League over the last 30 seasons. Construct a frequency distribution using 5 classes. Draw a histogram, a frequency polygon, and an ogive for the data, using relative frequencies. Describe the shape of the histogram.

40	43	40
53	47	46
44	57	43
43	52	44
54	47	51
39	48	36
37	56	42
54	56	49
54	52	40
48	50	40

Source: World Almanac and Book of Facts.

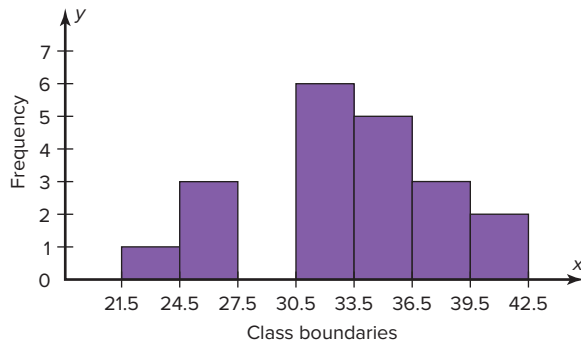
- 18. Protein Grams in Fast Food** The amount of protein (in grams) for a variety of fast-food sandwiches is reported here. Construct a frequency distribution, using 6 classes. Draw a histogram, a frequency polygon, and an ogive for the data, using relative frequencies. Describe the shape of the histogram.

23	30	20	27	44	26	35	20	29	29
25	15	18	27	19	22	12	26	34	15
27	35	26	43	35	14	24	12	23	31
40	35	38	57	22	42	24	21	27	33

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter.

Extending the Concepts

19. Using the histogram shown here, do the following.



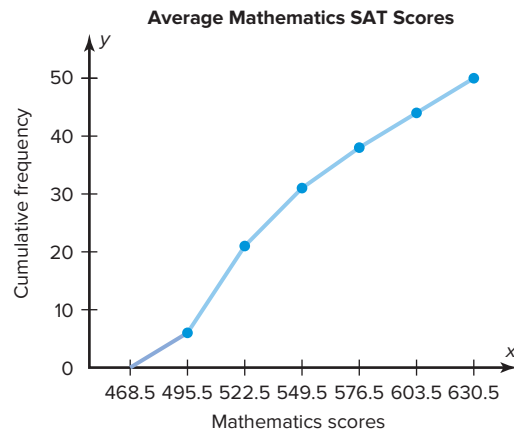
- Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies.
- Construct a frequency polygon.
- Construct an ogive.

20. Using the results from Exercise 19, answer these questions.

- How many values are in the class 27.5–30.5?
- How many values fall between 24.5 and 36.5?

- How many values are below 33.5?
- How many values are above 30.5?

21. **Math SAT Scores** Shown is an ogive depicting the cumulative frequency of the average mathematics SAT scores by state. Use it to construct a histogram and a frequency polygon.



Technology

TI-84 Plus Step by Step

Input

```
WINDOW
Xmin=100
Xmax=135
Xscl=5
Ymin=-5
Ymax=20
Yscl=5
Xres=1
```

Input

```
2nd [F1] Plot2 Plot3
Off
Type: [L1] [L2] [L3] [L4] [L5] [L6] [L7] [L8] [L9] [L0]
Xlist:L1
Freq:1
```

Step by Step

Constructing a Histogram

To display the graphs on the screen, enter the appropriate values in the calculator, using the **WINDOW** menu. The default values are $X_{\min} = -10$, $X_{\max} = 10$, $Y_{\min} = -10$, and $Y_{\max} = 10$.

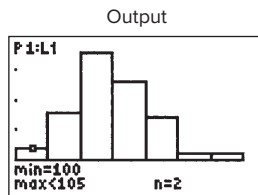
The X_{scl} changes the distance between the tick marks on the x axis and can be used to change the class width for the histogram.

To change the values in the **WINDOW**:

- Press **WINDOW**.
- Move the cursor to the value that needs to be changed. Then type in the desired value and press **ENTER**.
- Continue until all values are appropriate.
- Press **[2nd] [QUIT]** to leave the **WINDOW** menu.

To plot the histogram from raw data:

- Enter the data in L_1 .
- Make sure **WINDOW** values are appropriate for the histogram.
- Press **[2nd] [STAT PLOT] ENTER**.
- Press **ENTER** to turn the plot 1 on, if necessary.
- Move cursor to the Histogram symbol and press **ENTER**, if necessary. The histogram is the third option.
- Make sure **Xlist** is L_1 .
- Make sure **Freq** is 1.
- Press **GRAPH** to display the histogram.
- To obtain the frequency (number of data values in each class), press the **TRACE** key, followed by **◀** or **▶** keys.

**Example TI2–1**

Plot a histogram for the following data from Example 2–2.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Press **TRACE** and use the arrow keys to determine the number of values in each group.

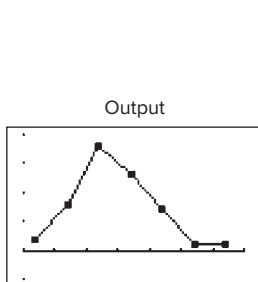
To graph a histogram from grouped data:

1. Enter the midpoints into L_1 .
2. Enter the frequencies into L_2 .
3. Make sure **WINDOW** values are appropriate for the histogram.
4. Press **[2nd] [STAT PLOT] ENTER**.
5. Press **ENTER** to turn the plot on, if necessary.
6. Move cursor to the histogram symbol, and press **ENTER**, if necessary.
7. Make sure **Xlist** is L_1 .
8. Make sure **Freq** is L_2 .
9. Press **GRAPH** to display the histogram.

Example TI2–2

Plot a histogram for the data from Examples 2–4 and 2–5.

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1



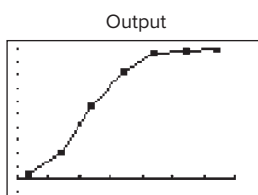
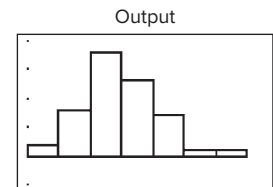
Input

L_1	L_2	L_3	1
102	2		
107	8		
112	18		
117	13		
122	7		
127	1		
132	1		

$L_1(1)=102$

Input

Plot1	Plot2	Plot3
On	Off	Off
Type:		
Xlist:	L_1	
Freq:	L_2	



To graph a frequency polygon from grouped data, follow the same steps as for the histogram except change the graph type from histogram (third graph) to a line graph (second graph).

To graph an ogive from grouped data, modify the procedure for the histogram as follows:

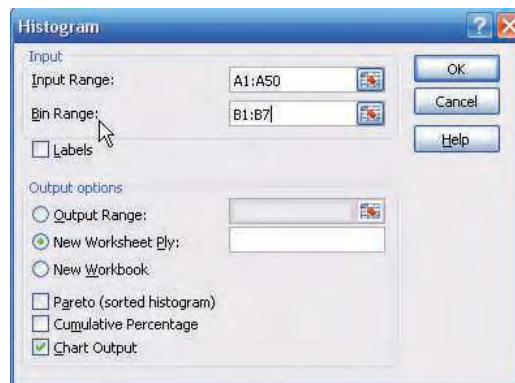
1. Enter the upper class boundaries into L_1 .
2. Enter the cumulative frequencies into L_2 .
3. Change the graph type from histogram (third graph) to line (second graph).
4. Change the $Y_{\max}$ from the **WINDOW** menu to the sample size.

EXCEL

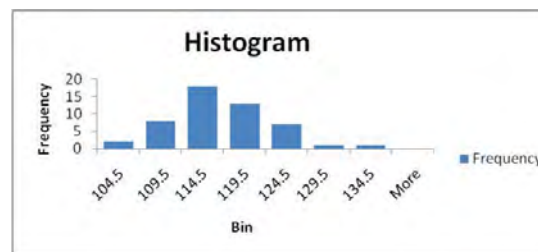
Step by Step

Constructing a Histogram

1. Press [Ctrl]-N for a new workbook.
2. Enter the data from Example 2–2 in column A, one number per cell.
3. Enter the upper boundaries into column B.
4. From the toolbar, select the Data tab, then select Data Analysis.
5. In Data Analysis, select Histogram and click [OK].
6. In the Histogram dialog box, type **A1:A50** in the Input Range box and type **B1:B7** in the Bin Range box.



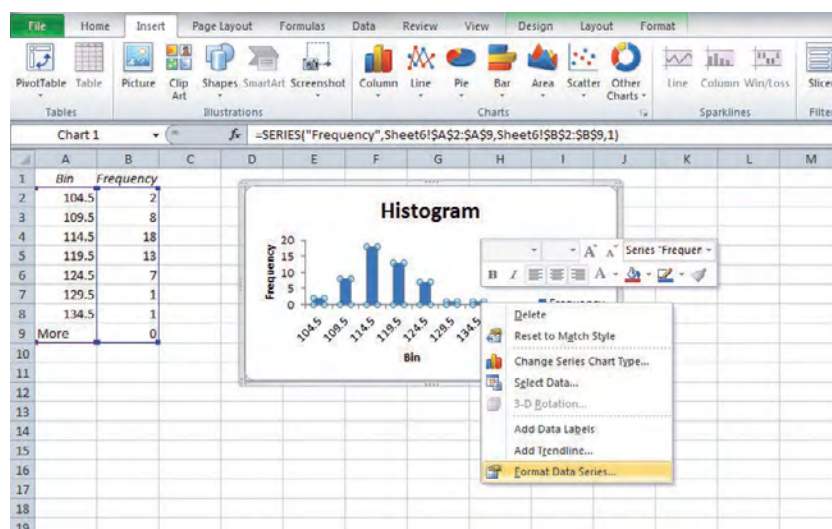
7. Select New Worksheet Ply and Chart Output. Click [OK].



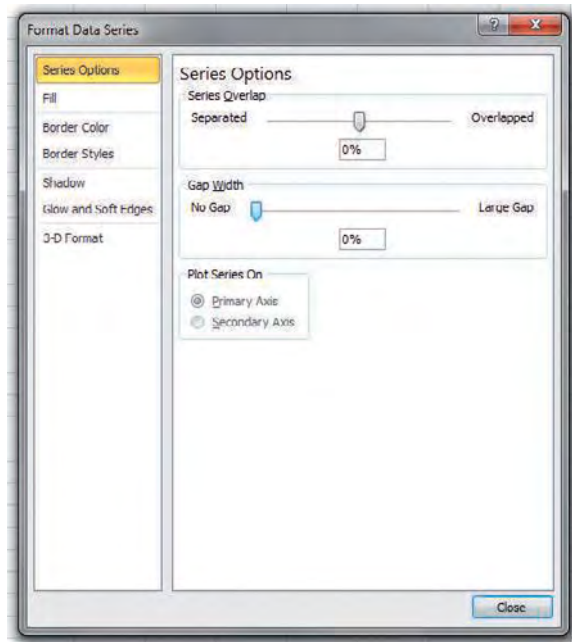
Editing the Histogram

To move the vertical bars of the histogram closer together:

1. Right-click one of the bars of the histogram, and select Format Data Series.

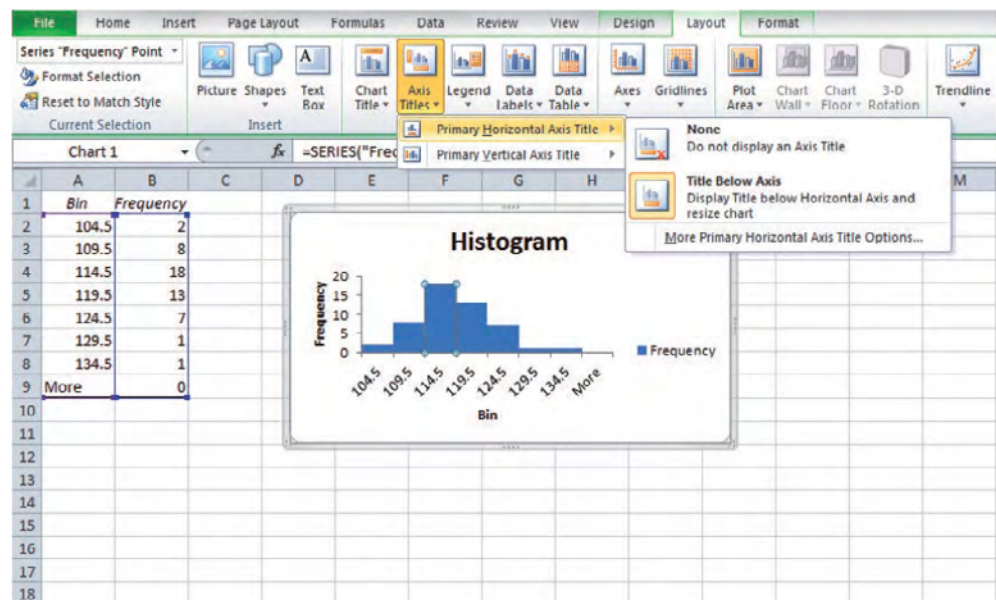


2. Move the Gap Width slider all the way to the left to change the gap width of the bars in the histogram to 0.



To change the label for the horizontal axis:

1. Left-click the mouse over any part of the histogram.
2. Select the Chart Tools tab from the toolbar.
3. Select the Layout tab, Axis Titles and Primary Horizontal Axis Title.



Once the Axis Titles text box is selected, you can type in the name of the variable represented on the horizontal axis.

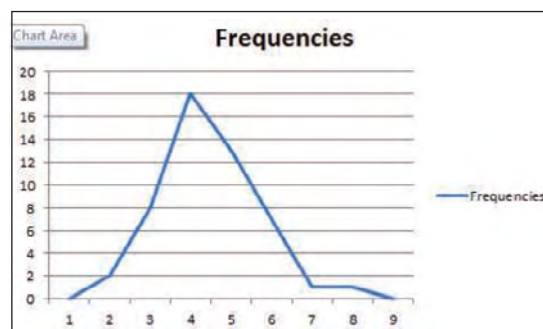
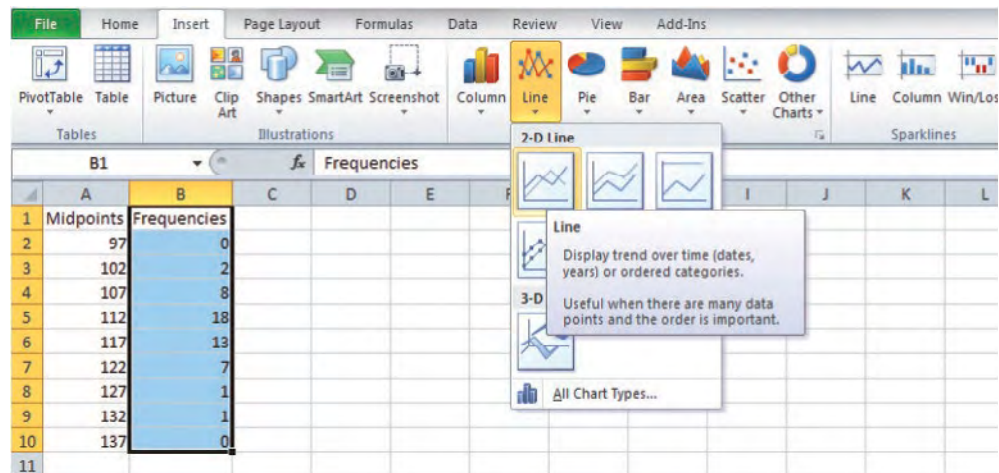
Constructing a Frequency Polygon

1. Press [CTRL]-N for a new notebook.
2. Enter the midpoints of the data from Example 2–2 into column A and the frequencies into column B, including labels.

Note: Classes with frequency 0 have been added at the beginning and the end to “anchor” the frequency polygon to the horizontal axis.

	A	B
1	Midpoints	Frequencies
2	97	0
3	102	2
4	107	8
5	112	18
6	117	13
7	122	7
8	127	1
9	132	1
10	137	0
11		

3. Press and hold the left mouse button, and drag over the Frequencies (including the label) from column B.
4. Select the Insert tab from the toolbar and the Line Chart option.
5. Select the 2-D line chart type.



We will need to edit the graph so that the midpoints are on the horizontal axis.

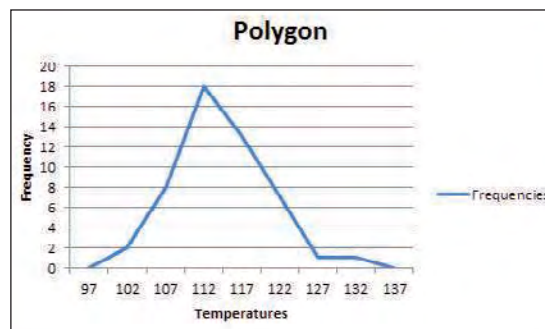
1. Right click the mouse on any region of the chart.
2. Choose Select Data.
3. Select Edit below the Horizontal (Category) Axis Labels panel on the right.
4. Press and hold the left mouse button, and drag over the midpoints (not including the label) for the Axis label range, then click [OK].
5. Click [OK] on the Select Data Source box.

Inserting Labels on the Axes

1. Click the mouse on any region of the graph.
2. Select Chart Tools and then Layout on the toolbar.
3. Select Axis Titles to open the horizontal and vertical axis text boxes. Then manually type in labels for the axes.

Changing the Title

1. Select Chart Tools, Layout from the toolbar.
2. Select Chart Title.
3. Choose one of the options from the Chart Title menu and edit.

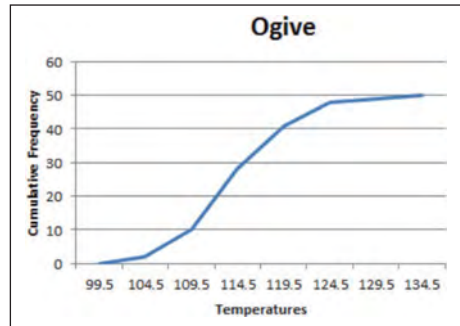


Constructing an Ogive

To create an ogive, use the upper class boundaries (horizontal axis) and cumulative frequencies (vertical axis) from the frequency distribution.

1. Type the upper class boundaries (including a class with frequency 0 before the lowest class to anchor the graph to the horizontal axis) and corresponding cumulative frequencies into adjacent columns of an Excel worksheet.
2. Press and hold the left mouse button, and drag over the Cumulative Frequencies from column B.
3. Select Line Chart, then the 2-D Line option.

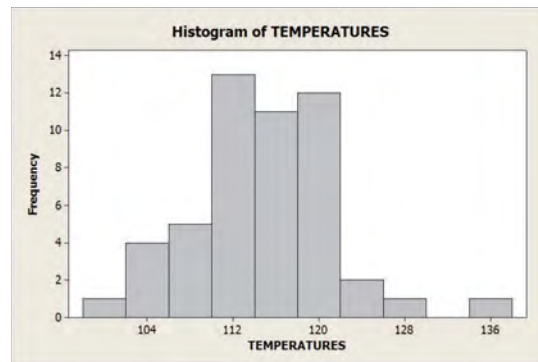
As with the frequency polygon, you can insert labels on the axes and a chart title for the ogive.



MINITAB Step by Step

Construct a Histogram

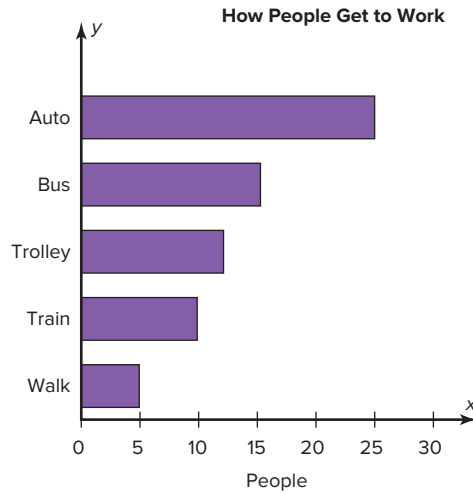
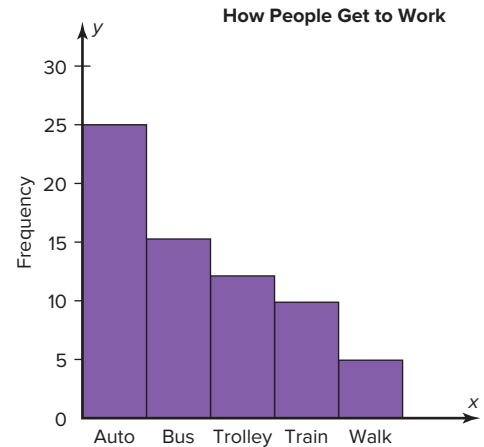
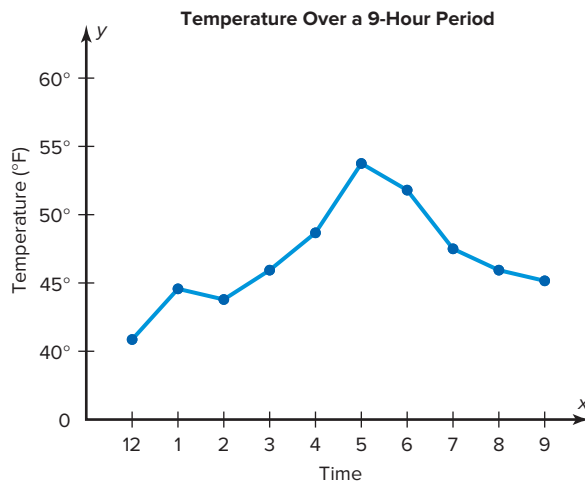
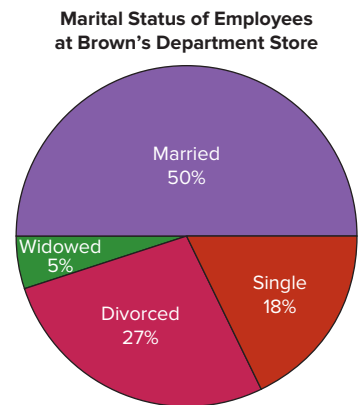
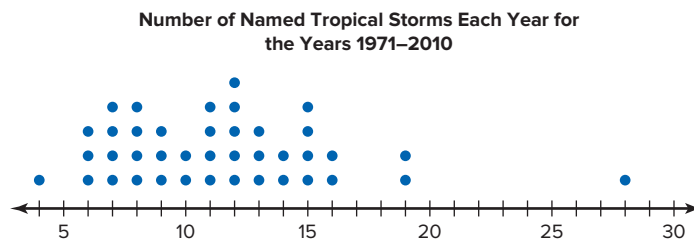
1. Enter the data from Example 2–2, the high temperatures for the 50 states, into C1.
2. Select **Graph>Histogram**.
3. Select [Simple], then click [OK].
4. Click C1 TEMPERATURES in the Graph variables dialog box.
5. Click [OK]. A new graph window containing the histogram will open.
6. Click the **File** menu to print or save the graph.



7. Click **File>Exit**.
8. Save the project as **Ch2-3.mpj**.

2–3 Other Types of Graphs

In addition to the histogram, the frequency polygon, and the ogive, several other types of graphs are often used in statistics. They are the bar graph, Pareto chart, time series graph, pie graph, and the dotplot. Figure 2–8 shows an example of each type of graph.

FIGURE 2-8 Other Types of Graphs Used in Statistics**(a)** Bar graph**(b)** Pareto chart**(c)** Time series graph**(d)** Pie graph**(e)** Dotplot**OBJECTIVE 3****Bar Graphs**

Represent data using bar graphs, Pareto charts, time series graphs, pie graphs, and dotplots.

When the data are qualitative or categorical, bar graphs can be used to represent the data. A bar graph can be drawn using either horizontal or vertical bars.

A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

EXAMPLE 2-8 College Spending for First-Year Students

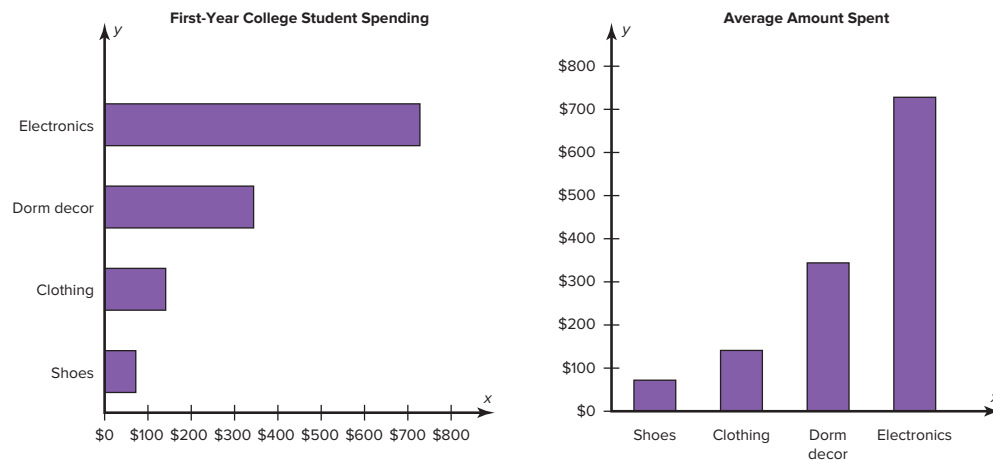
The table shows the average money spent by first-year college students. Draw a horizontal and vertical bar graph for the data.

Electronics	\$728
Dorm decor	344
Clothing	141
Shoes	72

Source: The National Retail Federation.

SOLUTION

1. Draw and label the x and y axes. For the horizontal bar graph place the frequency scale on the x axis, and for the vertical bar graph place the frequency scale on the y axis.
2. Draw the bars corresponding to the frequencies. See Figure 2-9.

FIGURE 2-9 Bar Graphs for Example 2-8

The graphs show that first-year college students spend the most on electronic equipment.

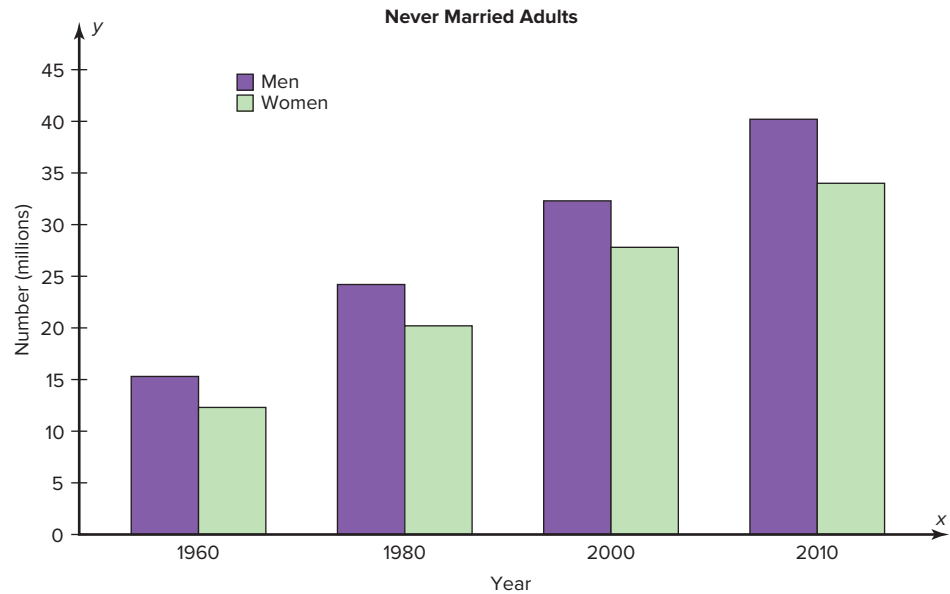
Bar graphs can also be used to compare data for two or more groups. These types of bar graphs are called *compound bar graphs*. Consider the following data for the number (in millions) of never married adults in the United States.

Year	Males	Females
1960	15.3	12.3
1980	24.2	20.2
2000	32.3	27.8
2010	40.2	34.0

Source: U.S. Census Bureau.

Figure 2-10 shows a bar graph that compares the number of never married males with the number of never married females for the years shown. The comparison is made by placing the bars next to each other for the specific years. The heights of the bars can be compared. This graph shows that there have consistently been more never married

FIGURE 2-10
Example of a Compound
Bar Graph



males than never married females and that the difference in the two groups has increased slightly over the last 50 years.

Pareto Charts

When the variable displayed on the horizontal axis is qualitative or categorical, a *Pareto chart* can also be used to represent the data.

A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

EXAMPLE 2-9 Traffic Congestion

The data shown consist of the average number of hours that a commuter spends in traffic congestion per year in each city. Draw and analyze a Pareto chart for the data.

City	Hours
Atlanta	52
Boston	64
Chicago	61
New York	74
Washington, D. C.	82

Source: 2015 Urban Mobility Scorecard

SOLUTION

Step 1 Arrange the data from the largest to the smallest according to the number of hours.

City	Hours
Washington, D.C.	82
New York	74
Boston	64
Chicago	61
Atlanta	52

Historical Note

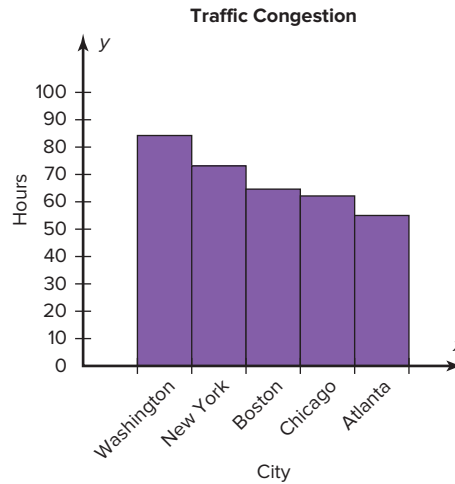
Vilfredo Pareto (1848–1923) was an Italian scholar who developed theories in economics, statistics, and the social sciences. His contributions to statistics include the development of a mathematical function used in economics. This function has many statistical applications and is called the Pareto distribution. In addition, he researched income distribution, and his findings became known as Pareto's law.

Step 2 Draw and label the x and y axes.

Step 3 Draw the vertical bars according to the number of hours (large to small).

The graph shown in Figure 2–11 shows that Washington, D.C. has the longest congestion times and Atlanta has the shortest times for the selection of cities.

FIGURE 2–11 Pareto Chart for Example 2–9



Suggestions for Drawing Pareto Charts

1. Make the bars the same width.
2. Arrange the data from largest to smallest according to frequency.
3. Make the units that are used for the frequency equal in size.

When you analyze a Pareto chart, make comparisons by looking at the heights of the bars.

The Time Series Graph

When data are collected over a period of time, they can be represented by a time series graph.

A **time series graph** represents data that occur over a specific period of time.

Example 2–10 shows the procedure for constructing a time series graph.

EXAMPLE 2–10 Price of an Advertisement for the Academy Awards Show

The data show the average cost (in millions of dollars) of a 30-second television ad on the Academy Awards show. Draw and analyze a time series graph for the data.

Year	2010	2011	2012	2013	2014	2015
Cost	1.40	1.55	1.61	1.65	1.78	1.90

Source: Kantar Media, USA TODAY RESEARCH

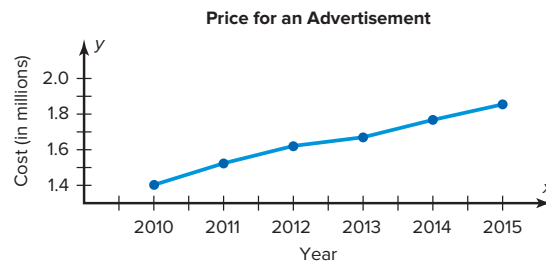
Historical Note

Time series graphs are over 1000 years old. The first ones were used to chart the movements of the planets and the sun.

SOLUTION

- Step 1** Draw and label the x and y axes.
- Step 2** Label the x axis for years and label the y axis for cost.
- Step 3** Plot each point for the values shown in the table.
- Step 4** Draw line segments connecting adjacent points. Do not try to fit a smooth curve through the data points. See Figure 2-12.

The data show that there has been an increase every year. The largest increase (shown by the steepest line segment) occurred for the year 2011 compared to 2010. The increases for the years 2011, 2012, and 2013 were relatively small compared to the increases from 2010 to 2014 and 2014 to 2015.

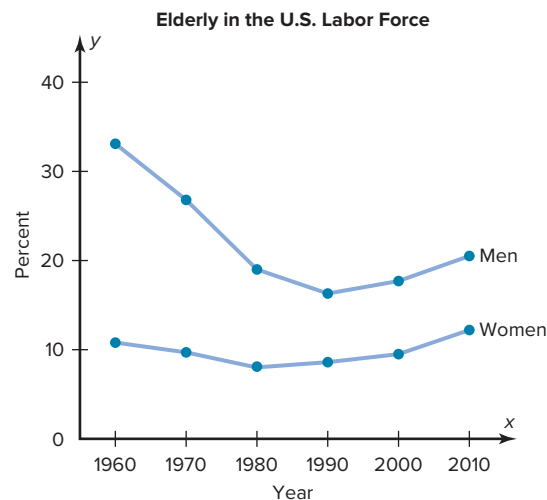
FIGURE 2-12 Figure for Example 2-10

When you analyze a time series graph, look for a trend or pattern that occurs over the time period. For example, is the line ascending (indicating an increase over time) or descending (indicating a decrease over time)? Another thing to look for is the slope, or steepness, of the line. A line that is steep over a specific time period indicates a rapid increase or decrease over that period.

Two or more data sets can be compared on the same graph called a *compound time series graph* if two or more lines are used, as shown in Figure 2-13. This graph shows

FIGURE 2-13

Two Time Series Graphs for Comparison



Source: Bureau of Census, U.S. Department of Commerce.

the percentage of elderly males and females in the U.S. labor force from 1960 to 2010. It shows that the percentage of elderly men decreased significantly from 1960 to 1990 and then increased slightly after that. For the elderly females, the percentage decreased slightly from 1960 to 1980 and then increased from 1980 to 2010.

The Pie Graph

Pie graphs are used extensively in statistics. The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections. Percentages or proportions can be used. The variable is nominal or categorical.

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Example 2–11 shows the procedure for constructing a pie graph.

EXAMPLE 2–11 Super Bowl Snack Foods

This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Pounds (frequency)
Potato chips	11.2 million
Tortilla chips	8.2 million
Pretzels	4.3 million
Popcorn	3.8 million
Snack nuts	2.5 million
	Total $n = 30.0$ million

Source: USA TODAY Weekend.

SOLUTION

Step 1 Since there are 360° in a circle, the frequency for each class must be converted to a proportional part of the circle. This conversion is done by using the formula

$$\text{Degrees} = \frac{f}{n} \cdot 360^\circ$$

where f = frequency for each class and n = sum of the frequencies. Hence, the following conversions are obtained. The degrees should sum to 360° .¹

Potato chips	$\frac{11.2}{30} \cdot 360^\circ = 134^\circ$
Tortilla chips	$\frac{8.2}{30} \cdot 360^\circ = 98^\circ$
Pretzels	$\frac{4.3}{30} \cdot 360^\circ = 52^\circ$
Popcorn	$\frac{3.8}{30} \cdot 360^\circ = 46^\circ$
Snack nuts	$\frac{2.5}{30} \cdot 360^\circ = 30^\circ$
Total	$\underline{\quad\quad\quad} 360^\circ$

¹Note: The degrees column does not always sum to 360° due to rounding.



The graph shows the number of murders (in thousands) that have occurred in the United States since 2001. Based on the graph, do you think the number of murders is increasing, decreasing, or remaining the same?



Source: Crime in the United States 2015, FBI, Department of Justice.

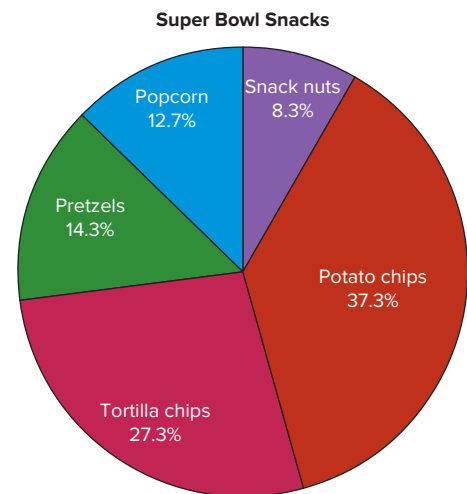
Step 2 Each frequency must also be converted to a percentage. Recall from Example 2–1 that this conversion is done by using the formula

$$\% = \frac{f}{n} \cdot 100$$

Hence, the following percentages are obtained. The percentages should sum to 100%.²

Potato chips	$\frac{11.2}{30} \cdot 100 = 37.3\%$
Tortilla chips	$\frac{8.2}{30} \cdot 100 = 27.3\%$
Pretzels	$\frac{4.3}{30} \cdot 100 = 14.3\%$
Popcorn	$\frac{3.8}{30} \cdot 100 = 12.7\%$
Snack nuts	$\frac{2.5}{30} \cdot 100 = 8.3\%$
Total	<u>99.9%</u>

FIGURE 2–14 Pie Graph for Example 2–11



Step 3 Next, using a protractor and a compass, draw the graph, using the appropriate degree measures found in Step 1, and label each section with the name and percentages, as shown in Figure 2–14.

²Note: The percent column does not always sum to 100% due to rounding.

EXAMPLE 2-12 Police Calls

Construct and analyze a pie graph for the calls received each shift by a local municipality for a recent year. (Data obtained by author.)

Shift	Frequency
1. Day	2594
2. Evening	2800
3. Night	2436
	7830

SOLUTION

Step 1 Find the number of degrees for each shift, using the formula:

$$\text{Degrees} = \frac{f}{n} \cdot 360^\circ$$

For each shift, the following results are obtained:

$$\text{Day: } \frac{2594}{7830} \cdot 360^\circ = 119^\circ$$

$$\text{Evening: } \frac{2800}{7830} \cdot 360^\circ = 129^\circ$$

$$\text{Night: } \frac{2436}{7830} \cdot 360^\circ = 112^\circ$$

Step 2 Find the percentages:

$$\text{Day: } \frac{2594}{7830} \cdot 100 = 33\%$$

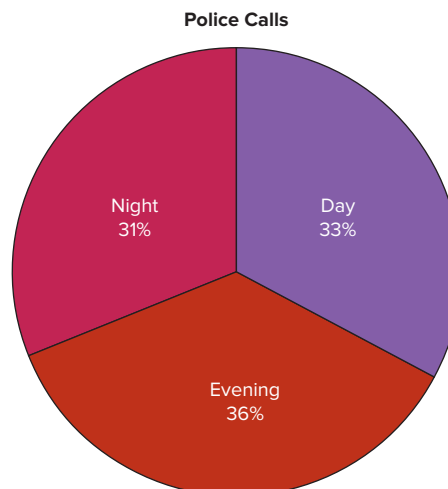
$$\text{Evening: } \frac{2800}{7830} \cdot 100 = 36\%$$

$$\text{Night: } \frac{2436}{7830} \cdot 100 = 31\%$$

Step 3 Using a protractor, graph each section and write its name and corresponding percentage as shown in Figure 2-15.

FIGURE 2-15

Figure for Example 2-12



To analyze the nature of the data shown in the pie graph, look at the size of the sections in the pie graph. For example, are any sections relatively large compared to the rest? Figure 2-15 shows that the number of calls for the three shifts are about equal, although slightly more calls were received on the evening shift.

Note: Computer programs can construct pie graphs easily, so the mathematics shown here would only be used if those programs were not available.

Dotplots

A dotplot uses points or dots to represent the data values. If the data values occur more than once, the corresponding points are plotted above one another.

A **dotplot** is a statistical graph in which each data value is plotted as a point (dot) above the horizontal axis.

Dotplots are used to show how the data values are distributed and to see if there are any extremely high or low data values.

EXAMPLE 2-13 Named Storms

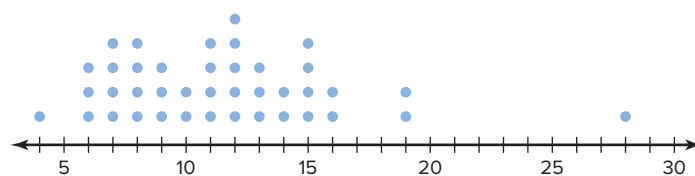
The data show the number of named storms each year for the last 40 years. Construct and analyze a dotplot for the data.

19	15	14	7	6	11	11
9	16	8	8	11	9	8
16	12	13	14	13	12	7
15	15	19	11	4	6	13
10	15	7	12	6	10	
28	12	8	7	12	9	

Source: NOAA.

- Step 1** Find the lowest and highest data values, and decide what scale to use on the horizontal axis. The lowest data value is 4 and the highest data value is 28, so a scale from 4 to 28 is needed.
- Step 2** Draw a horizontal line, and draw the scale on the line.
- Step 3** Plot each data value above the line. If the value occurs more than once, plot the other point above the first point. See Figure 2-16.

FIGURE 2-16 Figure for Example 2-13



The graph shows that the majority of the named storms occur with frequency between 6 and 16 per year. There are only 3 years when there were 19 or more named storms per year.

Stem and Leaf Plots

The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing. It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.

OBJECTIVE 4

Draw and interpret a stem and leaf plot.

A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

For example, a data value of 34 would have 3 as the stem and 4 as the leaf. A data value of 356 would have 35 as the stem and 6 as the leaf.

Example 2–14 shows the procedure for constructing a stem and leaf plot.

EXAMPLE 2–14 Out Patient Cardiograms

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

SOLUTION

Step 1 Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32,
32, 33, 36, 43, 44, 44, 45, 51, 52, 57

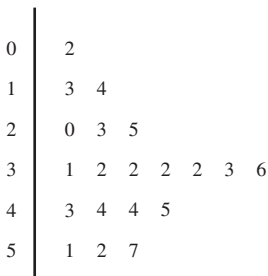
Note: Arranging the data in order is not essential and can be cumbersome when the data set is large; however, it is helpful in constructing a stem and leaf plot. The leaves in the final stem and leaf plot should be arranged in order.

Step 2 Separate the data according to the first digit, as shown.

02	13, 14	20, 23, 25	31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45	51, 52, 57		

Step 3 A display can be made by using the leading digit as the *stem* and the trailing digit as the *leaf*. For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf. For the value 14, the 1 is the stem and the 4 is the leaf. Now a plot can be constructed as shown in Figure 2–17.

FIGURE 2–17
Stem and Leaf Plot for
Example 2–14



Leading digit (stem)	Trailing digit (leaf)
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Figure 2–17 shows that the distribution peaks in the center and that there are no gaps in the data. For 7 of the 20 days, the number of patients receiving cardiograms was between 31 and 36. The plot also shows that the testing center treated from a minimum of 2 patients to a maximum of 57 patients in any one day.

If there are no data values in a class, you should write the stem number and leave the leaf row blank. Do not put a zero in the leaf row.

How Much Paper Money Is in Circulation Today?

The Federal Reserve estimated that during a recent year, there were 22 billion bills in circulation. About 35% of them were \$1 bills, 3% were \$2 bills, 8% were \$5 bills, 7% were \$10 bills, 23% were \$20 bills, 5% were \$50 bills, and 19% were \$100 bills. It costs about 3¢ to print each bill.

The average life of a \$1 bill is 22 months, a \$10 bill 3 years, a \$20 bill 4 years, a \$50 bill 9 years, and a \$100 bill 9 years. What type of graph would you use to represent the average lifetimes of the bills?



© Art Vandalay/Getty Images RF

EXAMPLE 2-15 Number of Car Thefts in a Large City

An insurance company researcher conducted a survey on the number of car thefts in a large city for a period of 30 days last summer. The raw data are shown. Construct a stem and leaf plot by using classes 50–54, 55–59, 60–64, 65–69, 70–74, and 75–79.

52	62	51	50	69
58	77	66	53	57
75	56	55	67	73
79	59	68	65	72
57	51	63	69	75
65	53	78	66	55

SOLUTION

Step 1 Arrange the data in order.

50, 51, 51, 52, 53, 53, 55, 55, 56, 57, 57, 58, 59, 62, 63, 65, 65, 66, 66, 67, 68, 69, 69, 72, 73, 75, 75, 77, 78, 79

Step 2 Separate the data according to the classes.

50, 51, 51, 52, 53, 53 55, 55, 56, 57, 57, 58, 59
62, 63 65, 65, 66, 66, 67, 68, 69, 69 72, 73
75, 75, 77, 78, 79

Step 3 Plot the data as shown here.

FIGURE 2-18

Stem and Leaf Plot for
Example 2-15

5	0	1	1	2	3	3		
5	5	5	6	7	7	8	9	
6	2	3						
6	5	5	6	6	7	8	9	9
7	2	3						
7	5	5	7	8	9			

Leading digit (stem)	Trailing digit (leaf)
5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

The graph for this plot is shown in Figure 2-18.

When you analyze a stem and leaf plot, look for peaks and gaps in the distribution. See if the distribution is symmetric or skewed. Check the variability of the data by looking at the spread.

Related distributions can be compared by using a back-to-back stem and leaf plot. The back-to-back stem and leaf plot uses the same digits for the stems of both distributions, but the digits that are used for the leaves are arranged in order out from the stems on both sides. Example 2–16 shows a back-to-back stem and leaf plot.

EXAMPLE 2–16 Number of Stories in Tall Buildings

The number of stories in two selected samples of tall buildings in Atlanta and Philadelphia is shown. Construct a back-to-back stem and leaf plot, and compare the distributions.

Atlanta					Philadelphia				
55	70	44	36	40	61	40	38	32	30
63	40	44	34	38	58	40	40	25	30
60	47	52	32	32	54	40	36	30	30
50	53	32	28	31	53	39	36	34	33
52	32	34	32	50	50	38	36	39	32
26	29								

Source: *The World Almanac and Book of Facts*.

SOLUTION

Step 1 Arrange the data for both data sets in order.

Step 2 Construct a stem and leaf plot, using the same digits as stems. Place the digits for the leaves for Atlanta on the left side of the stem and the digits for the leaves for Philadelphia on the right side, as shown. See Figure 2–19.

FIGURE 2–19 Back-to-Back Stem and Leaf Plot for Example 2–16

Atlanta		Philadelphia	
	9 8 6	2	5
8 6 4 4 2 2 2 2 1	3	0 0 0 0 2 2 3 4 6 6 6 8 8 9 9	
7 4 4 0 0	4	0 0 0 0	
5 3 2 2 0 0	5	0 3 4 8	
3 0	6	1	
0	7		

Step 3 Compare the distributions. The buildings in Atlanta have a large variation in the number of stories per building. Although both distributions are peaked in the 30- to 39-story class, Philadelphia has more buildings in this class. Atlanta has more buildings that have 40 or more stories than Philadelphia does.

Stem and leaf plots are part of the techniques called *exploratory data analysis*. More information on this topic is presented in Chapter 3.

Misleading Graphs

Graphs give a visual representation that enables readers to analyze and interpret data more easily than they could simply by looking at numbers. However, inappropriately drawn graphs can misrepresent the data and lead the reader to false conclusions. For example, a car manufacturer's ad stated that 98% of the vehicles it had sold in the past 10 years were still on the road. The ad then showed a graph similar to the one in Figure 2–20. The graph shows the percentage of the manufacturer's automobiles still on the road and the

percentage of its competitors' automobiles still on the road. Is there a large difference? Not necessarily.

Notice the scale on the vertical axis in Figure 2-20. It has been cut off (or truncated) and starts at 95%. When the graph is redrawn using a scale that goes from 0 to 100%, as in Figure 2-21, there is hardly a noticeable difference in the percentages. Thus, changing the units at the starting point on the y axis can convey a very different visual representation of the data.

FIGURE 2-20

Graph of Automaker's Claim Using a Scale from 95 to 100%

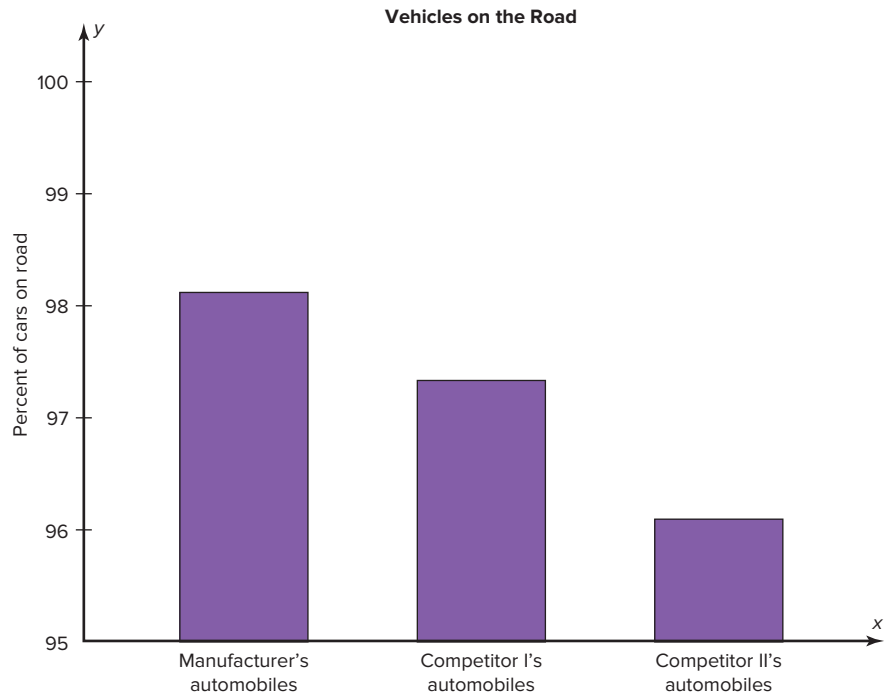
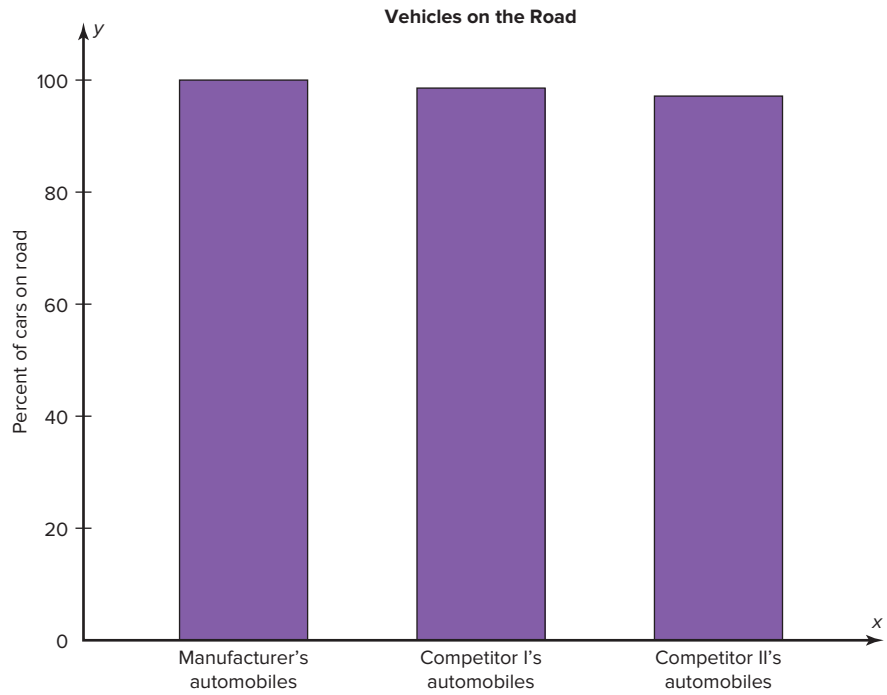


FIGURE 2-21

Graph in Figure 2-20 Redrawn Using a Scale from 0 to 100%



It is not wrong to truncate an axis of the graph; many times it is necessary to do so. However, the reader should be aware of this fact and interpret the graph accordingly. Do not be misled if an inappropriate impression is given.

Let us consider another example. The projected required fuel economy in miles per gallon for General Motors vehicles is shown. In this case, an increase from 21.9 to 23.2 miles per gallon is projected.

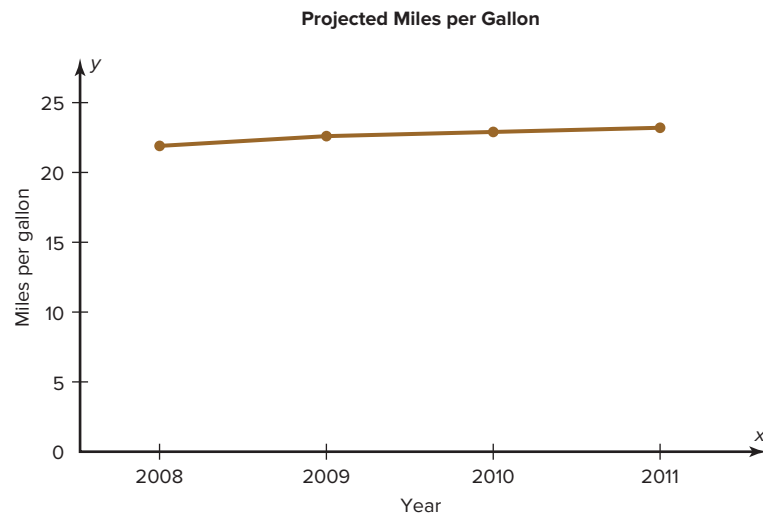
Year	2008	2009	2010	2011
MPG	21.9	22.6	22.9	23.2

Source: National Highway Traffic Safety Administration.

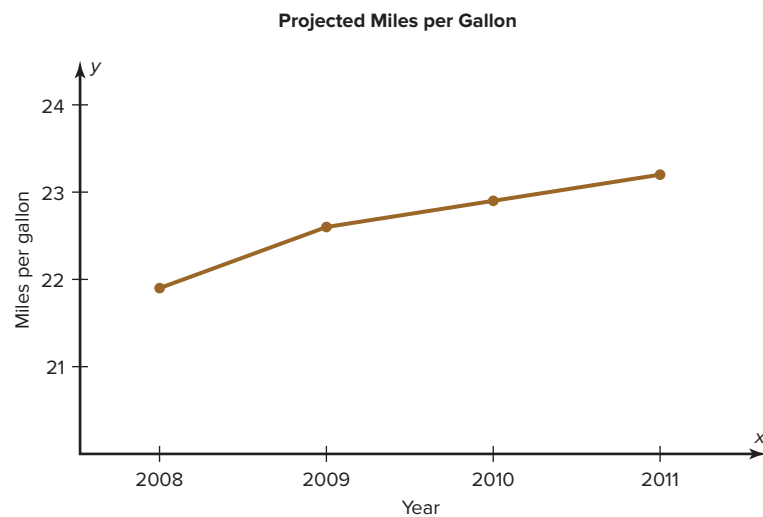
When you examine the graph shown in Figure 2–22(a), using a scale of 0 to 25 miles per gallon, the graph shows a slight increase. However, when the scale is changed to 21 to 24 miles per gallon, the graph shows a much larger increase even though the data remain the same. See Figure 2–22(b). Again, by changing the units or starting point on the y axis, one can change the visual representation.

FIGURE 2–22

Projected Miles per Gallon



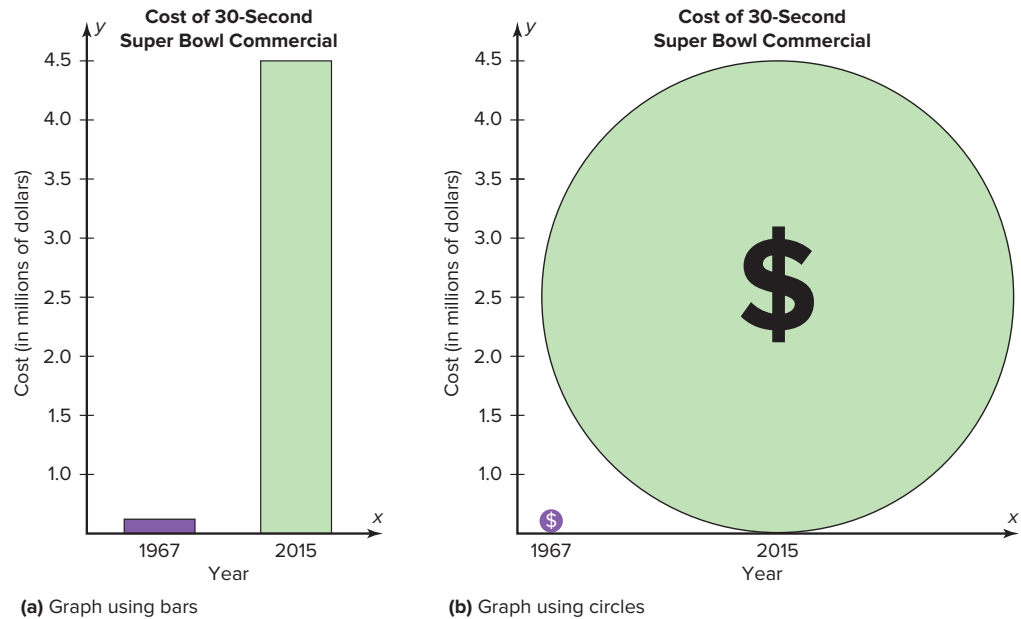
(a)



(b)

FIGURE 2-23

Comparison of Costs for
a 30-Second Super Bowl
Commercial



Another misleading graphing technique sometimes used involves exaggerating a one-dimensional increase by showing it in two dimensions. For example, the average cost of a 30-second Super Bowl commercial has increased from \$42,000 in 1967 to \$4.5 million in 2015 (*Source: USA TODAY*).

The increase shown by the graph in Figure 2-23(a) represents the change by a comparison of the heights of the two bars in one dimension. The same data are shown two-dimensionally with circles in Figure 2-23(b). Notice that the difference seems much larger because the eye is comparing the areas of the circles rather than the lengths of the diameters.

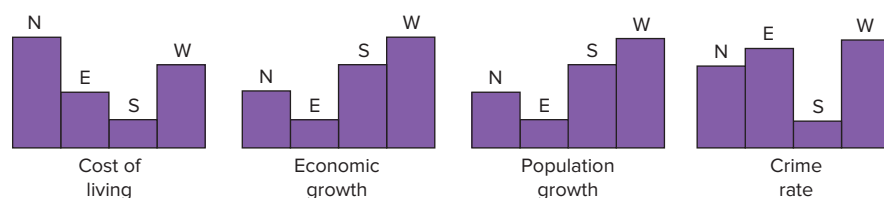
Note that it is not wrong to use the graphing techniques of truncating the scales or representing data by two-dimensional pictures. But when these techniques are used, the reader should be cautious of the conclusion drawn on the basis of the graphs.

Another way to misrepresent data on a graph is by omitting labels or units on the axes of the graph. The graph shown in Figure 2-24 compares the cost of living, economic growth, population growth, etc., of four main geographic areas in the United States. However, since there are no numbers on the y axis, very little information can be gained from this graph, except a crude ranking of each factor. There is no way to decide the actual magnitude of the differences.

Finally, all graphs should contain a source for the information presented. The inclusion of a source for the data will enable you to check the reliability of the organization presenting the data.

FIGURE 2-24

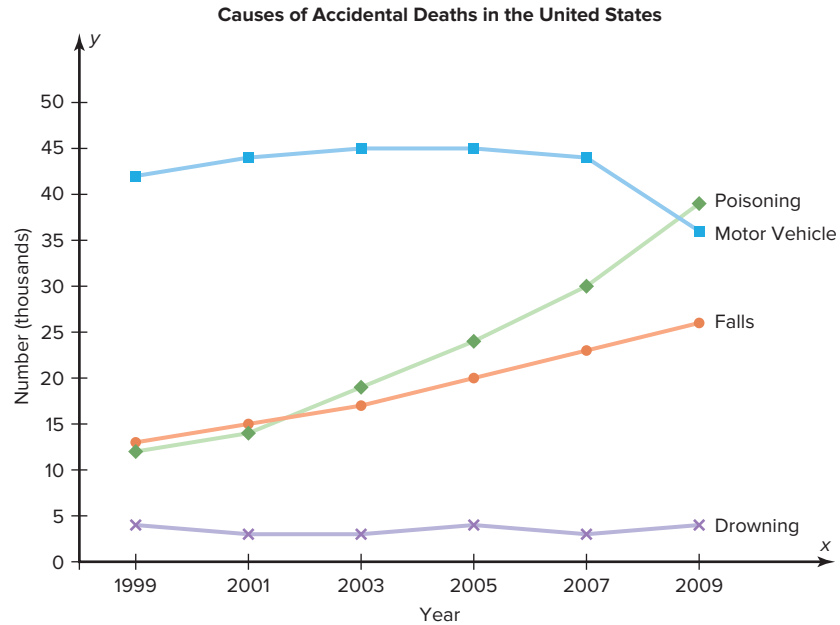
A Graph with No Units on
the y Axis



Applying the Concepts 2–3

Causes of Accidental Deaths in the United States, 1999–2009

The graph shows the number of deaths in the United States due to accidents. Answer the following questions about the graph.



Source: National Safety Council.

1. Name the variables used in the graph.
2. Are the variables qualitative or quantitative?
3. What type of graph is used here?
4. Which variable shows a decrease in the number of deaths over the years?
5. Which variable or variables show an increase in the number of deaths over the years?
6. The number of deaths in which variable remains about the same over the years?
7. List the approximate number of deaths for each category for the year 2001.
8. In 1999, which variable accounted for the most deaths? In 2009, which variable accounted for the most deaths?
9. In what year were the numbers of deaths from poisoning and falls about the same?

See page 108 for the answers.

Exercises 2–3

- 1. Tech Company Employees** Construct a vertical and horizontal bar graph for the number of employees (in thousands) of a sample of the largest tech companies as of 2014.

Company	Employees
IBM	380
Hewlett Packard	302
Xerox	147
Microsoft	128
Intel	107

Source: S & P Capital IQ

- 2. Worldwide Sales of Fast Foods** The worldwide sales (in billions of dollars) for several fast-food franchises for a specific year are shown. Construct a vertical bar graph and a horizontal bar graph for the data.

Wendy's	\$ 8.7
KFC	14.2
Pizza Hut	9.3
Burger King	12.7
Subway	10.0

Source: Franchise Times.

- 3. Gulf Coastlines** Construct a Pareto chart for the sizes of Gulf coastlines in statute miles for each state.

State	Coastline
Alabama	53
Florida	770
Louisiana	397
Mississippi	44
Texas	367

- 4. Roller Coaster Mania** The World Roller Coaster Census Report lists the following numbers of roller coasters on each continent. Represent the data graphically, using a Pareto chart.

Africa	17
Asia	315
Australia	22
Europe	413
North America	643
South America	45

Source: www.rcdb.com

- 5. Online Ad Spending** The amount spent (in billions of dollars) for ads online is shown. (The numbers for 2016 through 2019 are projected numbers.) Draw a time series graph and comment on the trend.

Year	2014	2015	2016	2017	2018	2019
Amount	\$19.72	\$31.53	\$43.83	\$53.29	\$61.14	\$69.04

Source: eMarketer.

- 6. Violent Crimes** The number of all violent crimes (murder, nonnegligent homicide, manslaughter, forcible rape, robbery, and aggravated assault) in the United States for each of these years is listed below. Represent the data with a time series graph.

2000	1,425,486	2004	1,360,088	2008	1,394,461
2001	1,439,480	2005	1,390,745	2009	1,325,896
2002	1,423,677	2006	1,435,123	2010	1,246,248
2003	1,383,676	2007	1,422,970		

Source: World Almanac and Book of Facts.

- 7. U.S. Licensed Drivers 70 or Older** Draw a time series graph for the number (in millions) of drivers in the United States 70 or older

Year	1982	1992	2002	2012
Number	10	15	20	23

Source: Federal Highway Administration

- 8. Valentine's Day Spending** The data show the average amount of money spent by consumers on Valentine's Day. Draw a time series graph for the data and comment on the trend.

Year	2007	2008	2009	2010	2011	2012
Amount	\$120	\$123	\$103	\$103	\$110	\$126

Source: National Retail Federation.

- 9. Credit Cards** Draw and analyze a pie graph for the number of credit cards a person has.

Number of cards	0	1	2 or 3	4 or more
Number	52	40	68	40

Source: Based on information from AARP Bulletin survey

- 10. Reasons We Travel** The following data are based on a survey from American Travel Survey on why people travel. Construct a pie graph for the data and analyze the results.

Purpose	Number
Personal business	146
Visit friends or relatives	330
Work-related	225
Leisure	299

Source: USA TODAY.

- 11. Kids and Guns** The following data show where children obtain guns for committing crimes. Draw and analyze a pie graph for the data.

Source	Friend	Family	Street	Gun or Pawn Shop	Other
Number	24	15	9	9	6

- 12. Colors of Automobiles** The popular car colors are shown. Construct a pie graph for the data.

White	19%
Silver	18
Black	16
Red	13
Blue	12
Gray	12
Other	10

Source: Dupont Automotive Color Popularity Report.

- 13. Ages of Football Players** The data show the ages of the players of the Super Bowl L Denver Bronco Champs in 2016. Construct a dotplot for the data, and comment on the distribution.

24	23	25	25	26	30
30	33	23	32	21	26
24	24	27	26	30	24
26	28	24	23	39	26
34	25	24	26	24	23
24	29	25	26	30	22
23	28	25	24	34	27
29	28	23	25	28	28
29	33	25	27	25	

Source: Fansided.com

- 14. Teacher Strikes** In Pennsylvania the numbers of teacher strikes for the last 14 years are shown. Construct a dotplot for the data. Comment on the graph.

9	13	15	7	7	14	9
10	14	18	7	8	8	3

Source: School Leader News.

- 15. Years of Experience** The data show the number of years of experience the players on the Pittsburgh Steelers football team have at the beginning of the season. Draw and analyze a dot plot for the data.

4 4 2 9 7 3 7 12 6
 5 1 4 5 2 7 6 12 3
 12 4 0 4 0 0 0 2 9
 2 6 7 13 4 2 6 9 4
 4 0 3 5 4 2 6 9 4
 4 0 3 5 3 11 1 4 2
 3 15 1 6 0 11 3 10 3

- 16. Commuting Times** Fifty off-campus students were asked how long it takes them to get to school. The times (in minutes) are shown. Construct a dotplot and analyze the data.

23 22 29 19 12
 18 17 30 11 27
 11 18 26 25 20
 25 15 24 21 31
 29 14 22 25 29
 24 12 30 27 21
 27 25 21 14 28
 17 17 24 20 26
 13 20 27 26 17
 18 25 21 33 29

- 17. 50 Home Run Club** There are 43 Major League baseball players (as of 2015) that have hit 50 or more home runs in one season. Construct a stem and leaf plot and analyze the data.

50 51 52 54 59 51 53
 54 50 58 51 54 53
 56 58 56 70 54 52
 58 54 64 52 73 57
 50 60 56 50 66 54
 52 51 58 63 57 52
 51 50 61 52 65 50

Source: *The World Almanac and Book of Facts*.

- 18. Calories in Salad Dressings** A listing of calories per 1 ounce of selected salad dressings (not fat-free) is given below. Construct a stem and leaf plot for the data.

100 130 130 130 110 110 120 130 140 100
 140 170 160 130 160 120 150 100 145 145
 145 115 120 100 120 160 140 120 180 100
 160 120 140 150 190 150 180 160

- 19. Length of Major Rivers** The data show the lengths (in hundreds of miles) of major rivers in South America and Europe. Construct a back-to-back stem and leaf plot, and compare the distributions.

South America				Europe			
39	21	10	10	5	12	7	6 8
11	10	2	10	5	5	4	6
10	14	10	12	18	5	13	9
17	15	10		14	6	6	11
15	25	16		8	6	3	4

Source: *The World Almanac and Book of Facts*.

- 20. Math and Reading Achievement Scores** The math and reading achievement scores from the National Assessment of Educational Progress for selected states are listed below. Construct a back-to-back stem and leaf plot with the data, and compare the distributions.

Math					Reading				
52	66	69	62	61	65	76	76	66	67
63	57	59	59	55	71	70	70	66	61
55	59	74	72	73	61	69	78	76	77
68	76	73			77	77	80		

Source: *World Almanac*.

- 21.** State which type of graph (Pareto chart, time series graph, or pie graph) would most appropriately represent the data.
- Situations that distract automobile drivers
 - Number of persons in an automobile used for getting to and from work each day
 - Amount of money spent for textbooks and supplies for one semester
 - Number of people killed by tornados in the United States each year for the last 10 years
 - The number of pets (dogs, cats, birds, fish, etc.) in the United States this year
 - The average amount of money that a person spent for his or her significant other for Christmas for the last 6 years
- 22.** State which graph (Pareto chart, time series graph, or pie graph) would most appropriately represent the given situation.
- The number of students enrolled at a local college for each year during the last 5 years
 - The budget for the student activities department at a certain college for a specific year
 - The means of transportation the students use to get to school
 - The percentage of votes each of the four candidates received in the last election
 - The record temperatures of a city for the last 30 years
 - The frequency of each type of crime committed in a city during the year
- 23. U.S. Health Dollar** The U.S. health dollar is spent as indicated below. Construct two different types of graphs to represent the data.

Government administration	9.7%
Nursing home care	5.5
Prescription drugs	10.1
Physician and clinical services	20.3
Hospital care	30.5
Other (OTC drugs, dental, etc.)	23.9

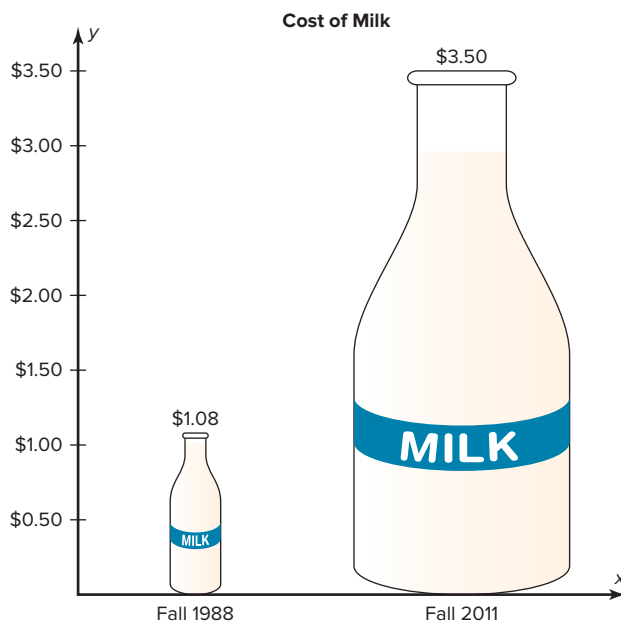
Source: *Time Almanac*.

- 24. Patents** The U.S. Department of Commerce reports the following number of U.S. patents received by foreign countries and the United States in the year 2010. Illustrate the data with a bar graph and a pie graph. Which do you think better illustrates this data set?

Japan	44,814	United Kingdom	4,302
Germany	12,363	China	2,657
South Korea	11,671	Israel	1,819
Taiwan	8,238	Italy	1,796
Canada	4,852	United States	107,792

Source: World Almanac.

- 25. Cost of Milk** The graph shows the increase in the price of a quart of milk. Why might the increase appear to be larger than it really is?



- 26. U.S. Population by Age** The following information was found in a recent almanac. Use a pie graph to illustrate the information. Is there anything wrong with the data?

U.S. Population by Age in 2011

Under 20 years	27.0%
20 years and over	73.0
65 years and over	13.1

Source: Time Almanac.

- 27. Chicago Homicides** Draw and compare two time series graphs for the number of homicides in the Chicago area.

Year	Homicides	As of June 29
2005	451	207
2007	448	204
2009	459	204
2011	435	187
2013	414	180

- 28. Trip Reimbursements** The average amount requested for business trip reimbursement is itemized below. Illustrate the data with an appropriate graph. Do you have any questions regarding the data?

Flight	\$440
Hotel stay	323
Entertainment	139
Phone usage	95
Transportation	65
Meal	38
Parking	34

Source: USA TODAY.

Technology

TI-84 Plus

Step by Step

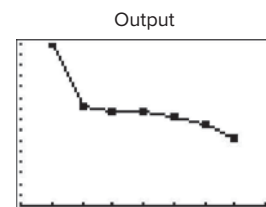
Step by Step

To graph a time series, follow the procedure for a frequency polygon from Section 2-2, using the following data for the number of outdoor drive-in theaters

Year	1988	1990	1992	1994	1996	1998	2000
Number	1497	910	870	859	826	750	637

Input			
L1	L2	L3	1
1988	1497	-----	
1990	910		
1992	870		
1994	859		
1996	826		
1998	750		
2000	637		
L1(1)=1988			

Input	
WINDOW	
Xmin=1986	
Xmax=2002	
Xscl=2	
Ymin=-5	
Ymax=1500	
Yscl=100	
Xres=1	



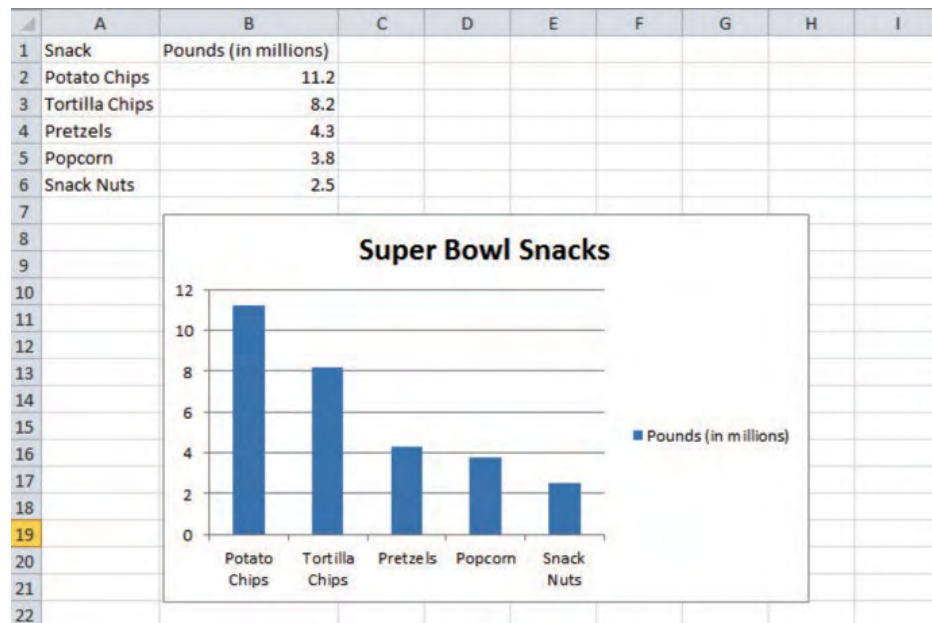
EXCEL

Step by Step

Constructing a Pareto Chart

To make a Pareto chart:

1. Enter the snack food categories from Example 2–11 into column A of a new worksheet.
2. Enter the corresponding frequencies in column B. The data should be entered in descending order according to frequency.
3. Highlight the data from columns A and B, and select the Insert tab from the toolbar.
4. Select the Column Chart type.
5. To change the title of the chart, click on the current title of the chart.
6. When the text box containing the title is highlighted, click the mouse in the text box and change the title.



Constructing a Time Series Chart

Example

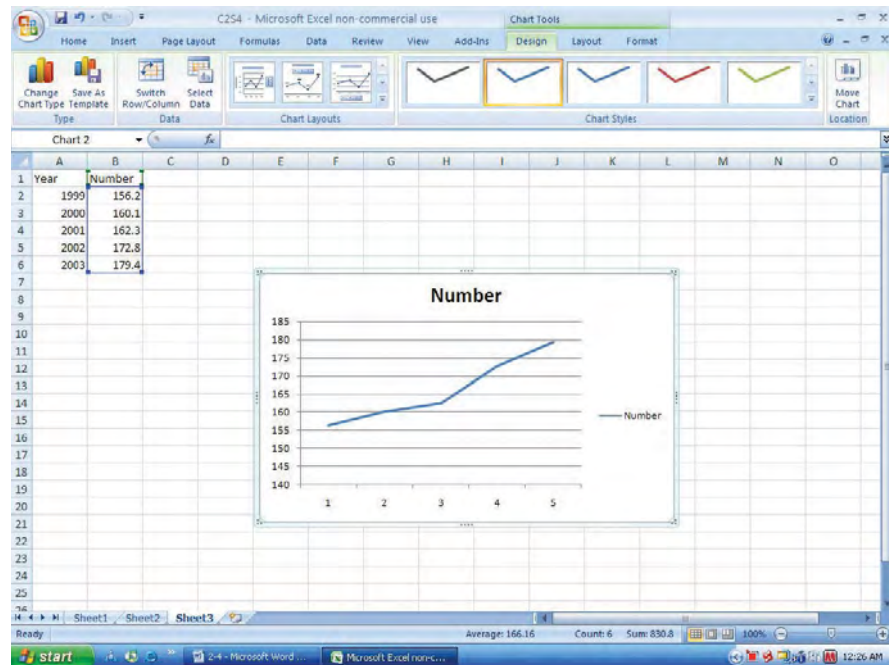
Year	1999	2000	2001	2002	2003
Vehicles*	156.2	160.1	162.3	172.8	179.4

*Vehicles (in millions) that used the Pennsylvania Turnpike.

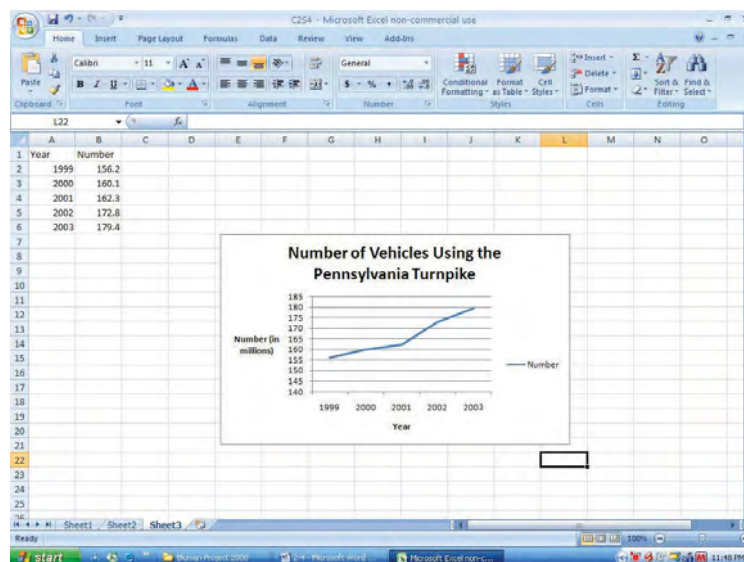
Source: Tribune Review.

To make a time series chart:

1. Enter the years 1999 through 2003 from the example in column A of a new worksheet.
2. Enter the corresponding frequencies in column B.
3. Highlight the data from column B and select the Insert tab from the toolbar.
4. Select the Line chart type.



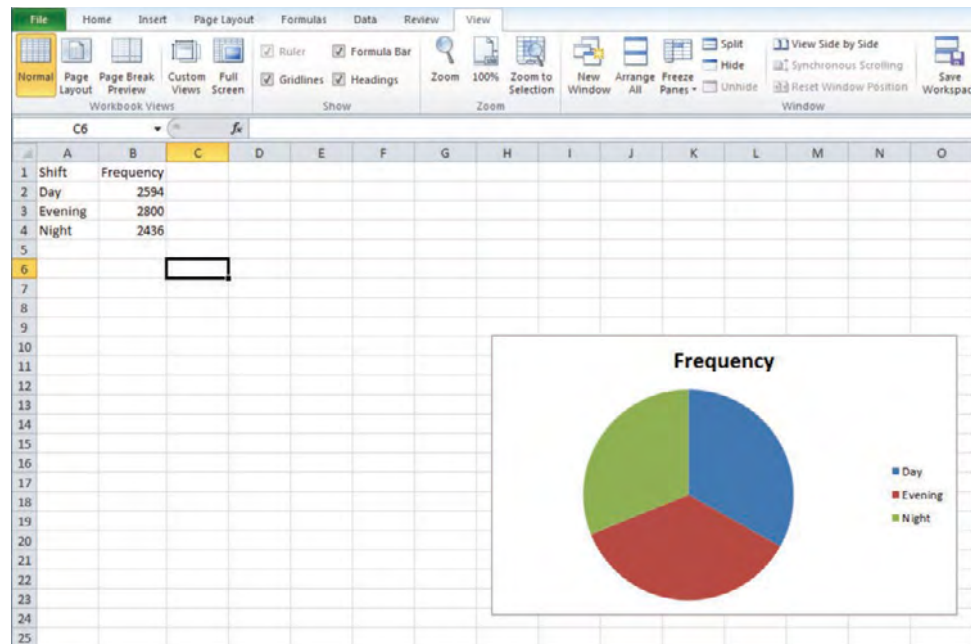
5. Right-click the mouse on any region of the graph.
6. Select the Select Data option.
7. Select Edit from the Horizontal Axis Labels and highlight the years from column A, then click [OK].
8. Click [OK] on the Select Data Source box.
9. Create a title for your chart, such as Number of Vehicles Using the Pennsylvania Turnpike Between 1999 and 2003. Right-click the mouse on any region of the chart. Select the Chart Tools tab from the toolbar, then Layout.
10. Select Chart Title and highlight the current title to change the title.
11. Select Axis Titles to change the horizontal and vertical axis labels.



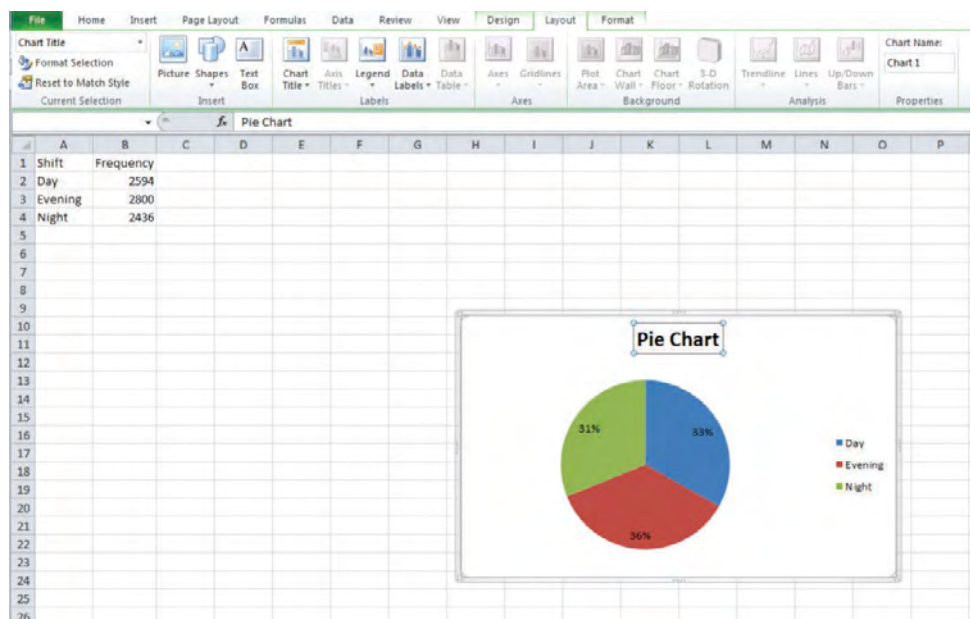
Constructing a Pie Chart

To make a pie chart:

1. Enter the shifts from Example 2–12 into column A of a new worksheet.
2. Enter the frequencies corresponding to each shift in column B.
3. Highlight the data in columns A and B and select Insert from the toolbar; then select the Pie chart type.



4. Click on any region of the chart. Then select Design from the Chart Tools tab on the toolbar.
5. Select Formulas from the chart Layouts tab on the toolbar.
6. To change the title of the chart, click on the current title of the chart.
7. When the text box containing the title is highlighted, click the mouse in the text box and change the title.



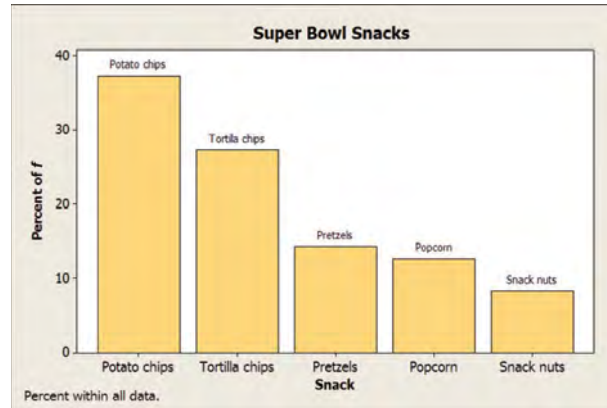
MINITAB

Step by Step

Construct a Bar Chart

The procedure for constructing a bar chart is similar to that for the pie chart.

1. Select **Graph>Bar Chart**.
 - a) Click on the drop-down list in Bars Represent: and then select values from a table.
 - b) Click on the Simple chart, then click [OK]. The dialog box will be similar to the Pie Chart Dialog Box.
2. Select the frequency column C2 f for Graph variables: and C1 Snack for the Categorical variable.



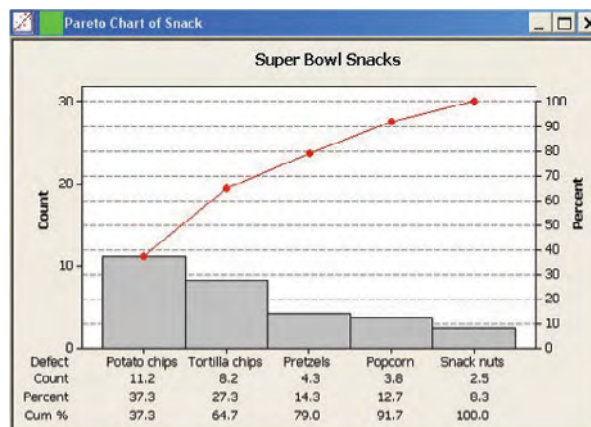
3. Click on [Labels], then type the title in the Titles/Footnote tab: **Super Bowl Snacks**.
4. Click the tab for Data Labels, then click the option to Use labels from column: and select C1 Snacks.
5. Click [OK] twice.

After the graph is made, right-click over any bar to change the appearance such as the color of the bars. To change the gap between them, right-click on the horizontal axis and then choose Edit X scale. In the Space Between Scale Categories select Gap between clusters then change the 1.5 to 0.2. Click [OK]. To change the y Scale to percents, right-click on the vertical axis and then choose Graph options and Show Y as a Percent.

Construct a Pareto Chart

Pareto charts are a quality control tool. They are similar to a bar chart with no gaps between the bars, and the bars are arranged by frequency.

1. Select **Stat>Quality Tools>Pareto**.
2. Click the option to Chart defects table.
3. Click in the box for the Labels in: and select C1 Snack.
4. Click on the frequencies column C2 f.



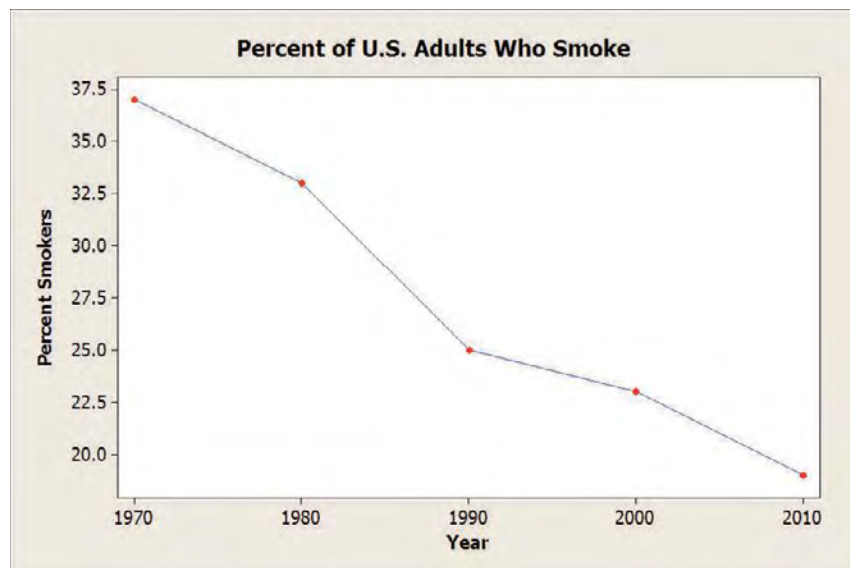
5. Click on [Options].
 - a) Type **Snack** for the X axis label and **Count** for the Y axis label.
 - b) Type in the title, **Super Bowl Snacks**.
6. Click [OK] twice. The chart is completed.

Construct a Time Series Plot

The data used are the percentage of U.S. adults who smoke (Example 2–10).

Year	1970	1980	1990	2000	2010
Number	37	33	25	23	19

1. Add a blank worksheet to the project by selecting **File>New>New-Minitab Worksheet**.
2. To enter the dates from 1970 to 2010 in C1, select **Calc>Make Patterned Data>Simple Set of Numbers**.
 - a) Type **Year** in the text box for Store patterned data in.
 - b) From First value: should be **1970**.
 - c) To Last value: should be **2010**.
 - d) In steps of should be **10** (for every 10-year increment). The last two boxes should be 1, the default value.
 - e) Click [OK]. The sequence from 1970 to 2010 will be entered in C1 whose label will be Year.
3. Type **Percent Smokers** for the label row above row 1 in C2.
4. Type **37** for the first number, then press [Enter].
5. Continue entering each value in a row of C2.

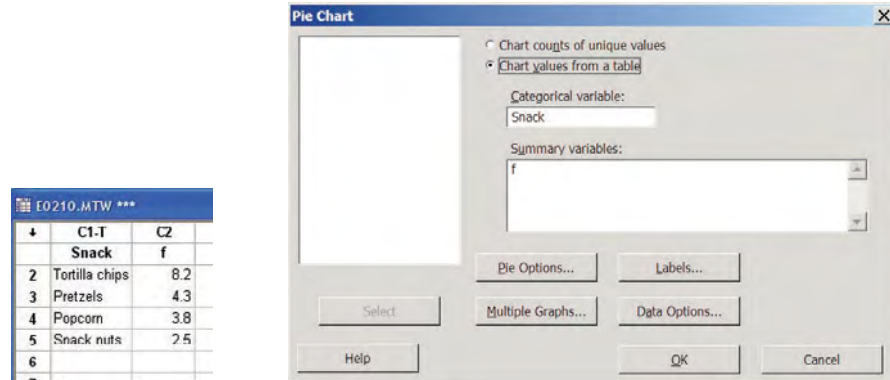


6. To make the graph, select **Graph>Time series plot**, then Simple, and press [OK].
 - a) For Series select Percent Smokers; then click [Time/scale].
 - b) Click the Stamp option and select Year for the Stamp column.
 - c) Click the Gridlines tab and select all three boxes, Y major, Y minor, and X major.

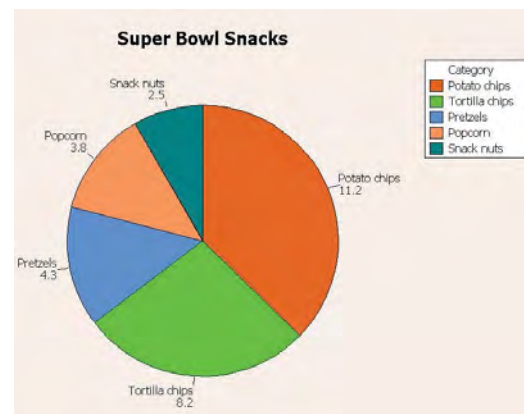
- d) Click [OK] twice. A new window will open that contains the graph.
- e) To change the title, double-click the title in the graph window. A dialog box will open, allowing you to change the text to Percent of U.S. Adults Who Smoke.

Construct a Pie Chart

1. Enter the summary data for snack foods and frequencies from Example 2–11 into C1 and C2.



2. Name them **Snack** and **f**.
3. Select **Graph>Pie Chart**.
 - a) Click the option for Chart summarized data.
 - b) Press [Tab] to move to Categorical variable, then double-click C1 to select it.
 - c) Press [Tab] to move to Summary variables, and select the column with the frequencies f.

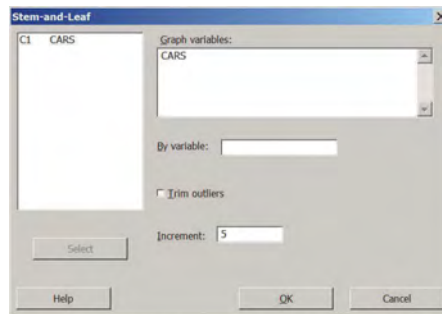


4. Click the [Labels] tab, then Titles/Footnotes.
 - a) Type in the title: **Super Bowl Snacks**.
 - b) Click the Slice Labels tab, then the options for Category name and Frequency.
 - c) Click the option to Draw a line from label to slice.
 - d) Click [OK] twice to create the chart.

Construct a Stem and Leaf Plot

1. Type in the data for Example 2–15. Label the column **CarThefts**.
2. Select **STAT>EDA>Stem-and-Leaf**. This is the same as **Graph>Stem-and-Leaf**.

3. Double-click on C1 CarThefts in the column list.
4. Click in the Increment text box, and enter the class width of 5.
5. Click [OK]. This character graph will be displayed in the session window.



Stem-and-Leaf Display: CarThefts

Stem-and-leaf of CarThefts N = 30
Leaf Unit = 1.0

6	5	011233
13	5	5567789
15	6	23
15	6	55667899
7	7	23
5	7	55789

Summary

- When data are collected, the values are called raw data. Since very little knowledge can be obtained from raw data, they must be organized in some meaningful way. A frequency distribution using classes is the common method that is used. (2–1)
- Once a frequency distribution is constructed, graphs can be drawn to give a visual representation of the data. The most commonly used graphs in statistics are the histogram, frequency polygon, and ogive. (2–2)
- Other graphs such as the bar graph, Pareto chart, time series graph, pie graph and dotplot can also be used. Some of these graphs are frequently seen in newspapers, magazines, and various statistical reports. (2–3)
- A stem and leaf plot uses part of the data values as stems and part of the data values as leaves. This graph has the advantage of a frequency distribution and a histogram. (2–3)
- Finally, graphs can be misleading if they are drawn improperly. For example, increases and decreases over time in time series graphs can be exaggerated by truncating the scale on the y axis. One-dimensional increases or decreases can be exaggerated by using two-dimensional figures. Finally, when labels or units are purposely omitted, there is no actual way to decide the magnitude of the differences between the categories. (2–3)

Important Terms

bar graph 75	cumulative frequency distribution 48	lower class limit 44	time series graph 78
categorical frequency distribution 43	dotplot 83	ogive 59	ungrouped frequency distribution 49
class 42	frequency 42	open-ended distribution 46	upper class limit 44
class boundaries 45	frequency distribution 42	Pareto chart 77	
class midpoint 45	frequency polygon 58	pie graph 80	
class width 45	grouped frequency distribution 44	raw data 42	
compound bar graphs 76	histogram 57	relative frequency graph 61	
cumulative frequency 59		stem and leaf plot 84	

Important Formulas

Formula for the percentage of values in each class:

$$\% = \frac{f}{n} \cdot 100$$

where

f = frequency of class

n = total number of values

Formula for the range:

$$R = \text{highest value} - \text{lowest value}$$

Formula for the class width:

$$\text{Class width} = \text{upper boundary} - \text{lower boundary}$$

Formula for the class midpoint:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Formula for the degrees for each section of a pie graph:

$$\text{Degrees} = \frac{f}{n} \cdot 360^\circ$$

Review Exercises

Section 2–1

- 1. How People Get Their News** The Brunswick Research Organization surveyed 50 randomly selected individuals and asked them the primary way they received the daily news. Their choices were via newspaper (N), television (T), radio (R), or Internet (I). Construct a categorical frequency distribution for the data and interpret the results.

N	N	T	T	T	I	R	R	I	T
I	N	R	R	I	N	N	I	T	N
I	R	T	T	T	T	N	R	R	I
R	R	I	N	T	R	T	I	I	T
T	I	N	T	T	I	R	N	R	T

- 2. Men's World Hockey Champions** The United States won the Men's World Hockey Championship in 1933 and 1960. Below are listed the world champions for the last 30 years. Use this information to construct a frequency distribution of the champions. What is the difficulty with these data?

1982	USSR	1999	Czech Republic
1983	USSR	2000	Czech Republic
1984	Not held	2001	Czech Republic
1985	Czechoslovakia	2002	Slovakia
1986	USSR	2003	Canada
1987	Sweden	2004	Canada
1988	Not held	2005	Czech Republic
1989	USSR	2006	Sweden
1990	Sweden	2007	Canada
1991	Sweden	2008	Russia
1992	Sweden	2009	Russia
1993	Russia	2010	Czech Republic
1994	Canada	2011	Finland
1995	Finland	2012	Russia
1996	Czech Republic	2013	Sweden
1997	Canada	2014	Russia
1998	Sweden	2015	Canada

Source: *Time Almanac*.

- 3. BUN Count** The blood urea nitrogen (BUN) count of 20 randomly selected patients is given here in milligrams per deciliter (mg/dl). Construct an ungrouped frequency distribution for the data.

17	18	13	14
12	17	11	20
13	18	19	17
14	16	17	12
16	15	19	22

- 4. Wind Speed** The data show the average wind speed for 36 days in a large city. Construct an ungrouped frequency distribution for the data.

8	15	9	8	9	10
8	10	14	9	8	8
12	9	8	8	14	9
9	13	13	10	12	9
13	8	11	11	9	8
9	13	9	8	8	10

- 5. Waterfall Heights** The data show the heights (in feet) of notable waterfalls in North America. Organize the data into a grouped frequency distribution using 6 classes. This data will be used for Exercises 7, 9, and 11.

90	420	300	194
640	68	268	276
620	76	165	833
370	53	132	600
594	70	308	
574	215	109	
317	850	212	
300	256	187	

Source: National Geographic Society

- 6. Ages of the Vice Presidents at the Time of Their Death** The ages at the time of death of those Vice Presidents of the United States who have passed away are listed below. Use the data to construct a frequency distribution.

Use 6 classes. The data for this exercise will be used for Exercises 8, 10, and 12.

90	83	80	73	70	51	68	79	70	71
72	74	67	54	81	66	62	63	68	57
66	96	78	55	60	66	57	71	60	85
76	98	77	88	78	81	64	66	77	93
70									

Source: World Almanac and Book of Facts.

Section 2-2

- Find the relative frequency for the frequency distribution for the data in Exercise 5.
- Find the relative frequency for the frequency distribution for the data in Exercise 6.
- Construct a histogram, frequency polygon, and ogive for the data in Exercise 5.
- Construct a histogram, frequency polygon, and ogive for the data in Exercise 6.
- Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 5.
- Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 6.

Section 2-3

- Non-Alcoholic Beverages** The data show the yearly consumption (in gallons) of popular non-alcoholic beverages. Draw a vertical and horizontal bar graph to represent the data.

Soft drinks	52
Water	34
Milk	26
Coffee	21

Source: U.S. Department of Agriculture

- Calories of Nuts** The data show the number of calories per ounce in selected types of nuts. Construct vertical and horizontal bar graphs for the data.

Types	Calories
Peanuts	160
Almonds	170
Macadamia	200
Pecans	190
Cashews	160

- Crime** The data show the percentage of the types of crimes commonly committed in the United States. Construct a Pareto chart for the data.

Theft	55%
Burglary	20%
Motor Vehicle Theft	11%
Assault	8%
Rape & Homicide	1%

Source: FBI

- Pet Care** The data (in billions of dollars) show the estimated amount of money spent on pet care in the United States. Construct a Pareto chart for the data.

Type of care	Amount spent
Veterinarian care	\$14
Supplies and medicine	11
Grooming and boarding	4
Animal purchases	2

Source: American Pet Products Association.

- Broadway Stage Engagements** The data show the number of new Broadway productions for the seasons. Construct and analyze a time series graph for the data.

Season	New Productions
2004	39
2005	39
2006	35
2007	39
2008	43
2009	39
2010	42
2011	41
2012	46
2013	44

Source: The Broadway League

- High School Dropout Rate** The data show the high school dropout rate for students for the years 2003 to 2013. Construct a time series graph and analyze the graph.

Year	Percent
2003	9.9
2004	10.3
2005	9.4
2006	9.3
2007	8.7
2008	8.0
2009	8.1
2010	7.4
2011	7.1
2012	6.6
2013	6.8

Source: U.S. Department of Commerce.

- Spending of College Freshmen** The average amounts spent by college freshmen for school items are shown. Construct a pie graph for the data.

Electronics/computers	\$728
Dorm items	344
Clothing	141
Shoes	72

Source: National Retail Federation.

- 20. Smart Phone Insurance** Construct and analyze a pie graph for the people who did or did not buy insurance for their smart phones at the time of purchase.

Response	Frequency
Yes	573
No	557
Don't Remember	166
Not offered any	211

Source: Based on information from Anderson Analytics

- 21. Peyton Manning's Colts Career** Peyton Manning played for the Indianapolis Colts for 14 years. (He did not play in 2011.) The data show the number of touchdowns he scored for the years 1998–2010. Construct a dotplot for the data and comment on the graph.

26	33	27	49	31	27	33
26	26	29	28	31	33	

Source: NFL.com

- 22. Songs on CDs** The data show the number of songs on each of 40 CDs from the author's collection. Construct a dotplot for the data and comment on the graph.

10	14	18	11
11	15	16	10
10	17	10	15
22	9	14	12
18	12	12	15
21	22	20	15
10	19	20	21
17	9	13	15
11	12	12	9
14	20	12	10

- 23. Weights of Football Players** A local football team has 30 players; the weight of each player is shown. Construct a stem and leaf plot for the data. Use stems 20__, 21__, 22__, etc.

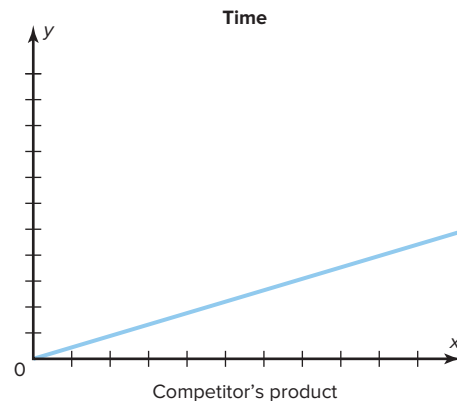
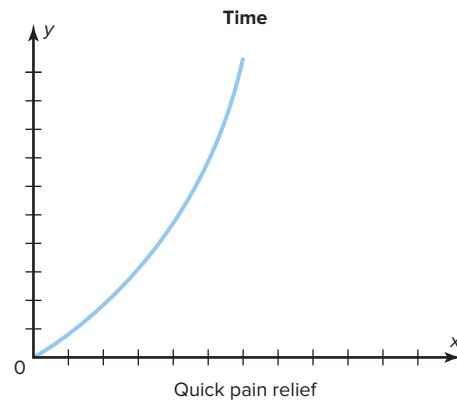
213	202	232	206	219
246	248	239	215	221
223	220	203	233	249
238	254	223	218	224
258	227	230	256	254
219	235	262	233	263

- 24. Public Libraries** The numbers of public libraries in operation for selected states are listed below. Organize the data with a stem and leaf plot.

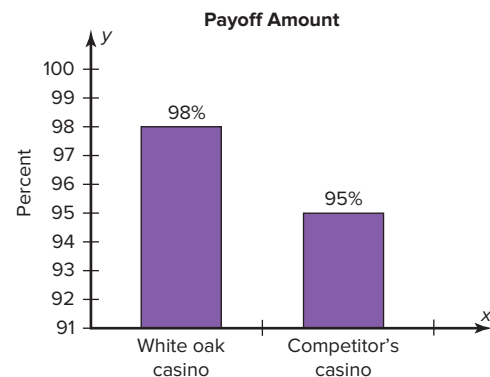
102	176	210	142	189	176	108	113	205
209	184	144	108	192	176			

Source: World Almanac.

- 25. Pain Relief** The graph below shows the time it takes Quick Pain Relief to relieve a person's pain. The graph below that shows the time a competitor's product takes to relieve pain. Why might these graphs be misleading?



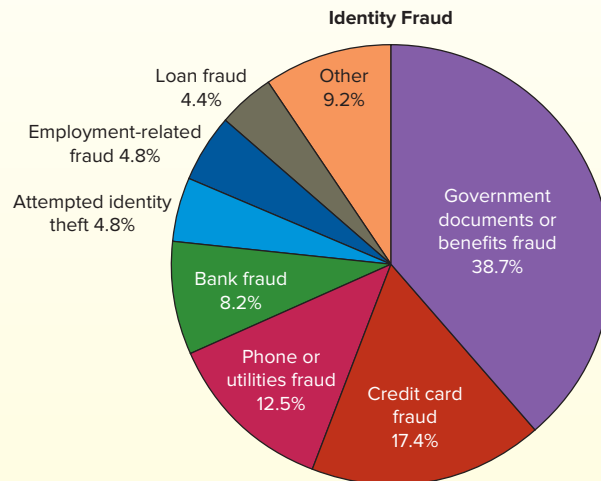
- 26. Casino Payoffs** The graph shows the payoffs obtained from the White Oak Casino compared to the nearest competitor's casino. Why is this graph misleading?



STATISTICS TODAY

How Your Identity Can Be Stolen —Revisited

Data presented in numerical form do not convey an easy-to-interpret conclusion; however, when data are presented in graphical form, readers can see the visual impact of the numbers. In the case of identity fraud, the reader can see that most of the identity frauds are due to lost or stolen wallets, checkbooks, or credit cards, and very few identity frauds are caused by online purchases or transactions.



The Federal Trade Commission suggests some ways to protect your identity:

1. Shred all financial documents no longer needed.
2. Protect your Social Security number.
3. Don't give out personal information on the phone, through the mail, or over the Internet.
4. Never click on links sent in unsolicited emails.
5. Don't use an obvious password for your computer documents.
6. Keep your personal information in a secure place at home.

Data Analysis

A Data Bank is found in Appendix B, or on the World Wide Web by following links from www.mhhe.com/math/stat/bluman

1. From the Data Bank located in Appendix B, choose one of the following variables: age, weight, cholesterol level, systolic pressure, IQ, or sodium level. Select at least 30 values. For these values, construct a grouped frequency distribution. Draw a histogram, frequency polygon, and ogive for the distribution. Describe briefly the shape of the distribution.
2. From the Data Bank, choose one of the following variables: educational level, smoking status, or exercise. Select at least 20 values. Construct an ungrouped frequency distribution for the data. For the distribution,

draw a Pareto chart and describe briefly the nature of the chart.

3. From the Data Bank, select at least 30 subjects and construct a categorical distribution for their marital status. Draw a pie graph and describe briefly the findings.
4. Using the data from Data Set IV in Appendix B, construct a frequency distribution and draw a histogram. Describe briefly the shape of the distribution of the tallest buildings in New York City.
5. Using the data from Data Set XI in Appendix B, construct a frequency distribution and draw a frequency polygon. Describe briefly the shape of the distribution for the number of pages in statistics books.

6. Using the data from Data Set IX in Appendix B, divide the United States into four regions, as follows:

Northeast CT ME MA NH NJ NY PA RI VT

Midwest IL IN IA KS MI MN MD MS NE ND OH SD WI

South AL AR DE DC FL GA KY LA MD NC OK SC TN TX VA WV

West AK AZ CA CO HI ID MT NV NM OR UT WA WY

Find the total population for each region, and draw a Pareto chart and a pie graph for the data. Analyze the results. Explain which chart might be a better representation for the data.

7. Using the data from Data Set I in Appendix B, make a stem and leaf plot for the record low temperatures in the United States. Describe the nature of the plot.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

- In the construction of a frequency distribution, it is a good idea to have overlapping class limits, such as 10–20, 20–30, 30–40.
- Bar graphs can be drawn by using vertical or horizontal bars.
- It is not important to keep the width of each class the same in a frequency distribution.
- Frequency distributions can aid the researcher in drawing charts and graphs.
- The type of graph used to represent data is determined by the type of data collected and by the researcher's purpose.
- In construction of a frequency polygon, the class limits are used for the x axis.
- Data collected over a period of time can be graphed by using a pie graph.

Select the best answer.

- What is another name for the ogive?
 - Histogram
 - Frequency polygon
 - Cumulative frequency graph
 - Pareto chart
- What are the boundaries for 8.6–8.8?
 - 8–9
 - 8.5–8.9
 - 8.55–8.85
 - 8.65–8.75
- What graph should be used to show the relationship between the parts and the whole?
 - Histogram
 - Pie graph
 - Pareto chart
 - Ogive

11. Except for rounding errors, relative frequencies should add up to what sum?

- 0
- 1
- 50
- 100

Complete these statements with the best answers.

- The three types of frequency distributions are _____, _____, and _____.
- In a frequency distribution, the number of classes should be between _____ and _____.
- Data such as blood types (A, B, AB, O) can be organized into a(n) _____ frequency distribution.
- Data collected over a period of time can be graphed using a(n) _____ graph.
- A statistical device used in exploratory data analysis that is a combination of a frequency distribution and a histogram is called a(n) _____.
- On a Pareto chart, the frequencies should be represented on the _____ axis.
- Housing Arrangements** A questionnaire on housing arrangements showed this information obtained from 25 respondents. Construct a frequency distribution for the data (H = house, A = apartment, M = mobile home, C = condominium). These data will be used in Exercise 19.

H	C	H	M	H	A	C	A	M
C	M	C	A	M	A	C	C	M
C	C	H	A	H	H	M		

19. Construct a pie graph for the data in Exercise 18.

20. **Items Purchased at a Convenience Store** When 30 randomly selected customers left a convenience store, each was asked the number of items he or she purchased. Construct an ungrouped frequency

distribution for the data. These data will be used in Exercise 21.

2	9	4	3	6
6	2	8	6	5
7	5	3	8	6
6	2	3	2	4
6	9	9	8	9
4	2	1	7	4

21. Construct a histogram, a frequency polygon, and an ogive for the data in Exercise 20.

22. **Coal Consumption** The following data represent the energy consumption of coal (in billions of Btu) by each of the 50 states and the District of Columbia. Use the data to construct a frequency distribution and a relative frequency distribution with 7 classes.

631	723	267	60	372	15	19	92	306	38
413	8	736	156	478	264	1015	329	679	1498
52	1365	142	423	365	350	445	776	1267	0
26	356	173	373	335	34	937	250	33	84
0	253	84	1224	743	582	2	33	0	426
474									

Source: Time Almanac.

23. Construct a histogram, frequency polygon, and ogive for the data in Exercise 22. Analyze the histogram.
24. **Recycled Trash** Construct a Pareto chart and a horizontal bar graph for the number of tons (in millions) of trash recycled per year by Americans based on an Environmental Protection Agency study.

Type	Amount
Paper	320.0
Iron/steel	292.0
Aluminum	276.0
Yard waste	242.4
Glass	196.0
Plastics	41.6

Source: USA TODAY.

25. **Identity Thefts** The results of a survey of 84 people whose identities were stolen using various methods are shown. Draw a pie chart for the information.

Lost or stolen wallet, checkbook, or credit card	38
Retail purchases or telephone transactions	15
Stolen mail	9
Computer viruses or hackers	8
Phishing	4
Other	10
	84

Source: Javelin Strategy and Research.

26. **Needless Deaths of Children** *The New England Journal of Medicine* predicted the number of needless deaths due to childhood obesity. Draw a time series graph for the data.

Year	2020	2025	2030	2035
Deaths	130	550	1500	3700

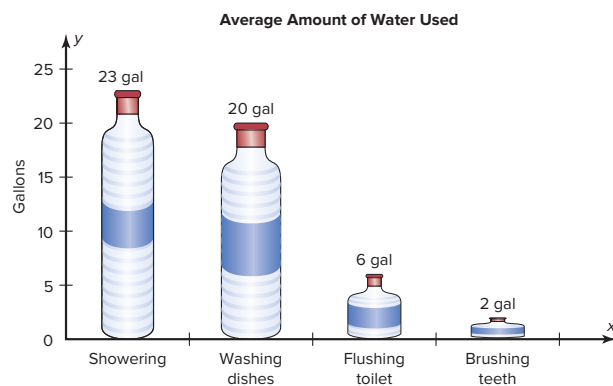
27. **Museum Visitors** The number of visitors to the Historic Museum for 25 randomly selected hours is shown. Construct a stem and leaf plot for the data.

15	53	48	19	38
86	63	98	79	38
62	89	67	39	26
28	35	54	88	76
31	47	53	41	68

28. **Parking Meter Revenue** In a small city the number of quarters collected from the parking meters is shown. Construct a dotplot for the data.

13	12	11	7	16
10	16	15	7	11
3	5	14	3	6
8	3	10	9	3
5	7	8	9	9
9	2	6	4	11
7	4	2	8	10
7	17	4	11	8
2	5	5	14	6
3	9	3	12	3

29. **Water Usage** The graph shows the average number of gallons of water a person uses for various activities. Can you see anything misleading about the way the graph is drawn?



Critical Thinking Challenges

1. The Great Lakes Shown are various statistics about the Great Lakes. Using appropriate graphs (your choice)

and summary statements, write a report analyzing the data.

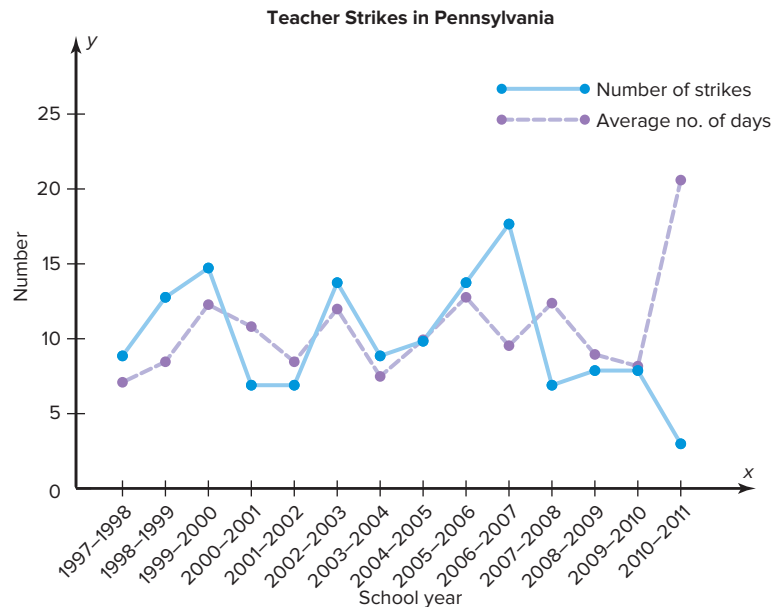
	Superior	Michigan	Huron	Erie	Ontario
Length (miles)	350	307	206	241	193
Breadth (miles)	160	118	183	57	53
Depth (feet)	1,330	923	750	210	802
Volume (cubic miles)	2,900	1,180	850	116	393
Area (square miles)	31,700	22,300	23,000	9,910	7,550
Shoreline (U.S., miles)	863	1,400	580	431	300

Source: *The World Almanac and Book of Facts*.

2. Teacher Strikes In Pennsylvania there were more teacher strikes in 2004 than there were in all other states combined. Because of the disruptions, state legislators want to pass a bill outlawing teacher strikes and submitting contract disputes to binding arbitration. The graph shows the number of teacher strikes in Pennsylvania for the school years 1997 to 2011. Use the graph to answer these questions.

- In what year did the largest number of strikes occur? How many were there?
- In what year did the smallest number of teacher strikes occur? How many were there?

- In what year was the average duration of the strikes the longest? How long was it?
- In what year was the average duration of the strikes the shortest? How long was it?
- In what year was the number of teacher strikes the same as the average duration of the strikes?
- Find the difference in the number of strikes for the school years 1997–1998 and 2010–2011.
- Do you think teacher strikes should be outlawed? Justify your conclusions.



Source: Pennsylvania School Boards Association.

Data Projects

Where appropriate, use the TI-84 Plus, Excel, MINITAB, or a computer program of your choice to complete the following exercises.

- Business and Finance** Consider the 30 stocks listed as the Dow Jones Industrials. For each, find its earnings

per share. Randomly select 30 stocks traded on the NASDAQ. For each, find its earnings per share. Create a frequency table with 5 categories for each data set. Sketch a histogram for each. How do the two data sets compare?

- Sports and Leisure** Use systematic sampling to create a sample of 25 National League and 25 American League baseball players from the most recently completed season. Find the number of home runs for each player. Create a frequency table with 5 categories for each data set. Sketch a histogram for each. How do the two leagues compare?
- Technology** Randomly select 50 songs from your music player or music organization program. Find the length (in seconds) for each song. Use these data to create a frequency table with 6 categories. Sketch a frequency polygon for the frequency table. Is the shape of the distribution of times uniform, skewed, or bell-shaped? Also note the genre of each song. Create a Pareto chart showing the frequencies of the various categories. Finally, note the year each song was released. Create a pie chart organized by decade to show the percentage of songs from various time periods.
- Health and Wellness** Use information from the Red Cross to create a pie chart depicting the percentages of Americans with various blood types. Also find information about blood donations and the percentage of each type donated. How do the charts compare? Why is the collection of type O blood so important?
- Politics and Economics** Consider the U.S. Electoral College System. For each of the 50 states, determine the number of delegates received. Create a frequency table with 8 classes. Is this distribution uniform, skewed, or bell-shaped?
- Your Class** Have each person in class take his or her pulse and determine the heart rate (beats in 1 minute). Use the data to create a frequency table with 6 classes. Then have everyone in the class do 25 jumping jacks and immediately take the pulse again after the activity. Create a frequency table for those data as well. Compare the two results. Are they similarly distributed? How does the range of scores compare?

Answers to Applying the Concepts

Section 2-1 Ages of Presidents at Inauguration

- The data were obtained from the population of all Presidents at the time this text was written.
- The oldest inauguration age was 69 years old.
- The youngest inauguration age was 42 years old.
- Answers will vary. One possible answer is

Age at inauguration	Frequency
42–45	2
46–49	7
50–53	8
54–57	16
58–61	5
62–65	4
66–69	2

- Answers will vary. For the frequency distribution given in question 4, there is a peak for the 54–57 class.
- Answers will vary. This frequency distribution shows no outliers. However, if we had split our frequency into 14 classes instead of 7, then the ages 42, 43, 68, and 69 might appear as outliers.
- Answers will vary. The data appear to be unimodal and fairly symmetric, centering on 55 years of age.

Section 2-2 Selling Real Estate

- A histogram of the data gives price ranges and the counts of homes in each price range. We can also talk about how the data are distributed by looking at a histogram.
- A frequency polygon shows increases or decreases in the number of home prices around values.

- A cumulative frequency polygon shows the number of homes sold at or below a given price.
- The house that sold for \$321,550 is an extreme value in this data set.
- Answers will vary. One possible answer is that the histogram displays the outlier well since there is a gap in the prices of the homes sold.
- The distribution of the data is skewed to the right.

Section 2-3 Causes of Accidental Deaths in the United States

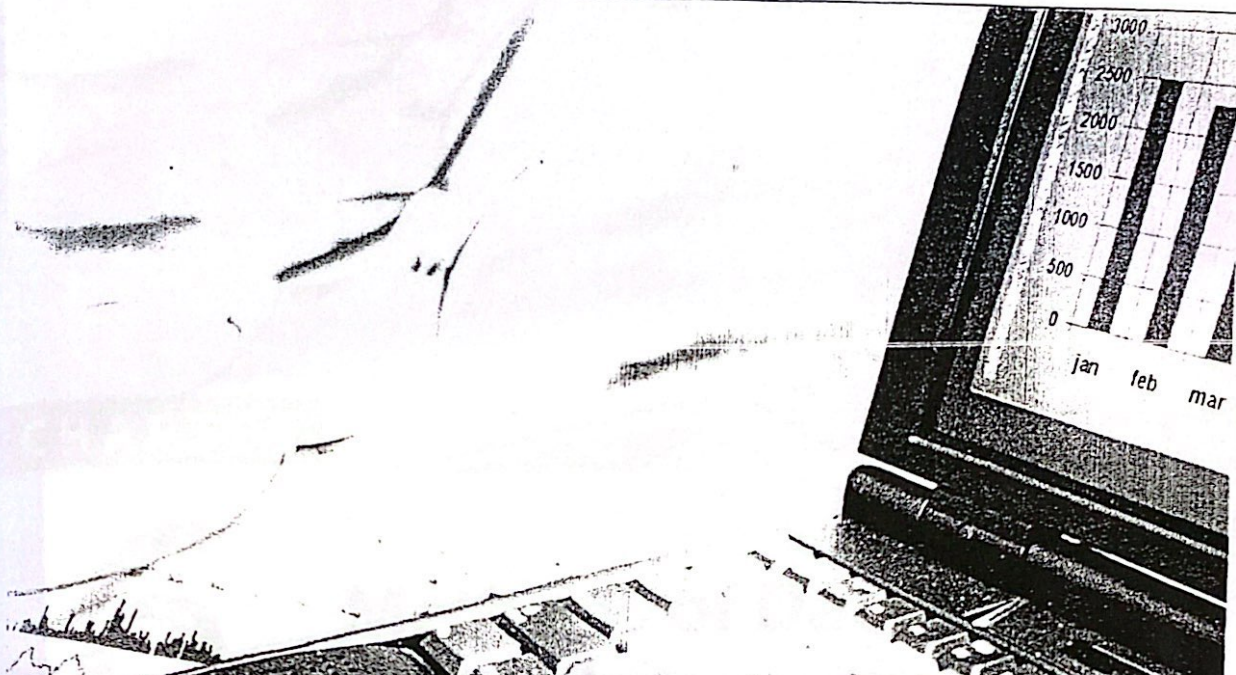
- The variables in the graph are the year, cause of death, and number of deaths in thousands.
- The cause of death is qualitative, while the year and number of deaths are quantitative.
- A time series graph is used here.
- The motor vehicle accidents showed a slight increase from 1999 to 2007, and then a decrease.
- The number of deaths due to poisoning and falls is increasing.
- The number of deaths due to drowning remains about the same over the years.
- For 2001, about 44,000 people died in motor vehicle accidents, about 14,000 people died from falls, about 15,000 people died from poisoning, and about 3000 people died from drowning.
- In 1999, motor vehicle accidents claimed the most lives, while in 2009, poisoning claimed the most lives.
- Around 2002, the number of deaths from falls and poisoning were about the same.

DATA COLLECTION METHODS

CONTENTS

Part III

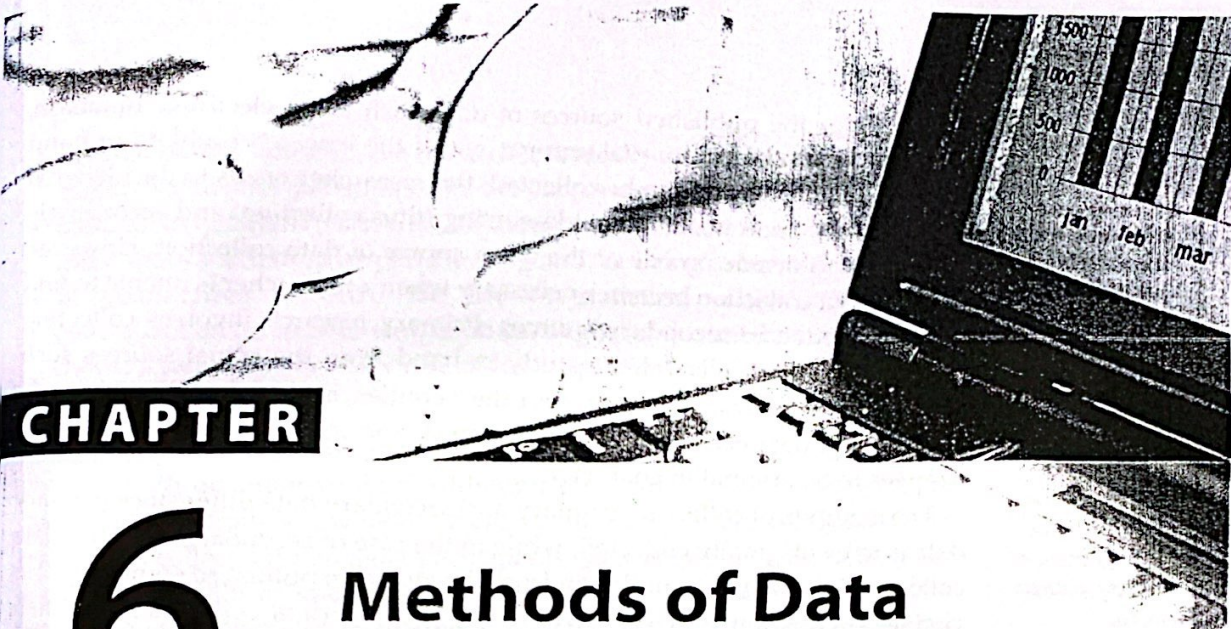
1. Chapter 6 Methods of Data Collection
2. Chapter 7 Measurement in Management and Scaling Techniques
3. Chapter 8 Questionnaire Design and Fieldwork Plan
4. Chapter 9 Sampling Design: Theory and Practice



PART III

Data Collection, Measurement and Sampling

- Chapter 6** Methods of Data Collection
- Chapter 7** Measurement in Management and Scaling Techniques
- Chapter 8** Questionnaire Design and Fieldwork Plan
- Chapter 9** Sampling Design: Theory and Practice
- Chapter 10** Experimental Design



CHAPTER

6

Methods of Data Collection

Learning Outcomes

After reading this chapter, student should be able to:

- Understand the difference between qualitative and quantitative methods of data collection
- Describe various types of data collection methods, and state their uses and limitations
- Use an appropriate method or a combination of different methods for data collection
- Identify ethical issues involved in business research and the ways of ensuring that research informants or subjects are not harmed by the study

6.0 Introduction

Once the research problem has been defined and research design has been set out for the study in hand, the task of data collection begins with the help of data collection methods. The data collection methods are the means to collect information about the objects under study in a systematic way. If data is collected haphazardly, it would be difficult to answer research questions in a conclusive way.

As discussed in Chapter 2, research can be divided into primary and secondary, based on the sources of data collection. Secondary research involves any information from published sources which has not been specifically collected for the current research problems.

■ The secondary research involves any information from published sources which has not been specifically collected for the current research problems.

■ The primary research involves collecting information specifically for the study in hand from the actual sources such as consumers, users/non-users or other entities involved in the research.

This includes the published sources of data such as an electronic database, periodicals, company's annual reports, etc. If the research problem in hand demands secondary data to be collected, the researcher needs to decide what kind of information he/she would be using (thus collecting) and accordingly he/she has to decide on one or the other source of data collection. However, primary data collection becomes necessary when a researcher is unable to find the data needed in secondary sources. Primary research involves collecting information specifically for the study in hand from the actual sources such as consumers, users/non-users or other entities involved in the research. The primary data is therefore collected fresh and for the first time, and thus happens to be original in character.

The methods of collecting primary and secondary data differ since primary data is to be originally collected, while in the case of secondary data, the data collection is simply a compilation from the available published source(s). This chapter focuses on different methods of primary data collection, and their merits and demerits.

6.1 Data Collection Method: Qualitative versus Quantitative

Broadly speaking, there are basically two methods of data collection—qualitative and quantitative. These two methods of data collection can be used in conjunction with each other.

■ The qualitative research explores attitudes, behaviour and experiences through methods such as interviews or focus groups.

Qualitative research explores attitudes, behaviour and experiences through methods such as interviews or focus groups. It attempts to get an in-depth opinion from participants. As qualitative research examines attitudes, behaviour and experiences which are related to the personal information of participants, fewer people take part in the research but the contact with these people tends to last a lot longer. Under the umbrella of qualitative research there are many different methods such as participants' observations, in-depth interview and focus group discussion.

■ Quantitative research generates statistics through the use of large-scale survey research, using methods such as questionnaires or structured interviews.

Quantitative research generates statistics through the use of large-scale survey research, using methods such as questionnaires or structured interviews. If a business analyst or market researcher has stopped you on the streets, or you have filled in a questionnaire which has arrived through the post, this falls under the umbrella of quantitative research. This type of research reaches many more people, but the contact with those people is much shorter than it is in qualitative research.

Following are the basic differences between the two:

- The typical qualitative data involves "words" and quantitative data involves "numbers".
- The qualitative research is inductive and quantitative research is deductive in nature. The qualitative research does not necessarily require setting up a hypothesis. However, development of the hypothesis is a must in quantitative research.

- In quantitative research, the researcher is ideally an objective observer. The researcher neither participates in the survey nor influences what is being studied. In qualitative research, on the other hand, the researcher learns the most about a situation by participating and/or being immersed in it.
- In quantitative research, the use of statistical analysis allows for generalization (to some extent) to others. A goal of quantitative research is to choose a sample that closely resembles the population. Qualitative research does not seek to choose samples that are representative of populations.

However, qualitative data does provide a depth and richness of data not possible with quantitative data. Although there are fewer participants, the researchers generally know more details about each participant. Quantitative researchers collect data on more participants, so it is not possible to have the depth and breadth of knowledge about each. It is often difficult to choose between quantitative and qualitative research design. At times, a researcher may choose a design because he or she is more familiar with one method or the other or a colleague recommends a particular design. However, the research will be more meaningful if the decision is made based on well-considered, suitable design rather than simply choosing a design that is more familiar or comfortable to the researcher. A choice between research methods rests fundamentally on a set of decisions about the questions a researcher wants to answer and the practicality of gathering the kind of data that will answer those questions. The first step is to look for an obvious fit.

In the following sub-sections, we shall discuss some of the important tools including both qualitative as well as quantitative methods, which are widely used by the researchers.

6.1.1 Observation

The observation method is the most commonly used method especially in studies relating to behavioural sciences. Observation is a technique that involves systematically selecting, watching and recording behaviour and characteristics of living beings, objects or phenomena. Under the observation method, the information is sought by way of investigator's own direct observations without the knowledge of the respondents. For example, in a study relating to consumer behaviour, the investigator instead of asking the brand of wrist watch used by respondents may look himself/herself at the watch. The main advantage of this method is that the subjective bias is eliminated as actual behaviours get recorded and thus, we get more accurate information. Secondly, the information obtained under this method relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes. Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents. This method is particularly suitable in studies which deal with subjects (respondents) who are not capable of giving verbal reports of their feelings for one reason or other. Here are some examples which will help to understand how observation method could be useful where other methods of data collection might not be suitable or fails.

■ Observation is a technique that involves systematically selecting, watching and recording behaviour and characteristics of living beings, objects or phenomena.

- 1 Suppose, the marketing manager of TESCO supermarket chain in Malaysia is interested to arrange the racks of items for kids (such as toys, chocolates, etc.). In order to increase the sales, the manager might be interested to arrange the most preferred and expensive items at the lower rack within the reach of kids so that they can pick the products of their interest and the intensity of force by kids to their parents to buy the products would be high. In this case observation could be the best method to identify what kind of toys and what flavour of chocolates are most preferred by kids, say below the age of 6 years. The method like survey may not be suitable for the kids who have just started picking up the language and not yet freely started interacting with unknown people.
- 2 The road traffic management in Malaysia is interested to know what the average waiting time is for a car near a particular toll gate during the peak hours, say 8 am to 10 am and 5 pm to 7 pm. In this case survey method will fail as most of the car owners (who often cross through that particular toll gate) may not be in a position to tell the exact waiting time. The best method is to fix up the camera for observation near the toll gate during the peak period and identify how many cars passed through that particular toll gate during the peak hours and thus the exact waiting time can be found out.
- 3 A video camera in a retail store can be used to record a customers' behaviour while she buys a garment. If it is a full-service store like many Indian markets (such as in Klang City in Malaysia), the customer could ask for a particular brand, look for specific colours, designs and fabrics, or prices and so on in a particular sequence. The customers' facial reactions or eagerness or lack of interest when a piece is displayed to them can be recorded along with the garments. The information collected through the video camera can later be used to identify the purchase factors, purchase behaviour, brand preference, price and colour preference, and matched with the customers' age and complexion.

While choosing observation as the method of data collection, the researcher should keep certain things in mind: What should be observed? How the observations should be recorded? Or how the accuracy of observations should be ensured? If the observation is characterized by careful definition of units to be observed, the style of recording the observed information, standardized conditions of observation and the selection of pertinent data of observation, then the observation is known as *structured observation*. On the other hand, if the observation is to take place without these characteristics to be sought in advance, it is known as *unstructured observation*. Structured observation is considered as appropriate in descriptive studies, whereas in an exploratory study, the observation is more likely to be unstructured.

However, the observational method has certain limitations. One of the major problem might be that we are not sure if a representative sample is chosen for the study, because recording of observational data take place at public places and we do not have control over who and how many are being observed at a given time.

6.1.2 Survey Method

A survey involves the collection of information from representative target respondents using a predesigned questionnaire. Unlike observation, it is structured method of data collection in which we extract exactly the same information from all the target population. There are basically four types of survey used by researchers which are described below.

■ The survey is a structured way to collect standardized information from individuals using a questionnaire.

Personal interview (face to face)

Among all the survey methods, personal interview (face-to-face) is the most widely used by researchers all over the world. These types of interviews consist of administering structured questionnaires and trained interviewers ask fixed, choice questions in a consistent format. In order to obtain a more accurate outcome of the overall survey, the interviewer should keep the following points in his/her mind. These important tips will ensure reliable, credible and unbiased responses from the respondents.

■ The personal interview consists of administering structured questionnaires where trained interviewers ask fixed, choice questions in a consistent format.

- The interviewer must be well-organized and knowledgeable in the subject matter presented.
- The interviewer should ask the same question without change in wording from all the respondents.
- The interviewer should ask every question in the same context.
- The interviewer should explain the purpose of the research to respondents in the same manner.
- The interviewer should ensure that each question is understood in the same way by all the respondents.
- The interviewer should write down the response in a standardized form to avoid any confusion.
- The interviewer should try to extract the unbiased response.
- The interviewer should be aware of the impact of his or her behaviour on the response of the respondents.

Advantages: Personal interview has the following advantages.

- If respondents find difficult to understand any question, the researcher can explain and clarify it to get the true and correct response on the question.
- Different measurement tools can be used in one survey depending on the requirement of the analysis by the researchers such as close-ended, visual open-ended questions, etc.
- The survey method has the advantage of getting the representative sample from the target population. So, findings of the research can be generalized. Response rate is generally very high.
- Extensive probing can be used to collect detailed information.
- Respondent's body language can guide the interviewer and be recorded to help interpret comments.

Mail survey

Imagine that you are interested in exploring the attitudes college students have about writing. Since it would be impossible to interview every student

on campus, choosing the mail-out survey as your method would enable you to choose a large sample of college students. You might choose to limit your research to your own college or university, or you might extend your survey to several different institutions. If your research question demands it, the mail survey allows you to sample a very broad group of subjects at a small cost.

Advantages: Mail survey has the following advantages.

- **Low cost:** Compared to a personal interview or self-administered survey, mail surveys cost less. The cost of mail survey could be 50% less than self-administered survey and 75% less than face-to-face survey.
- **Convenience:** As the survey is conducted through a mail/in process, the respondents are free to respond at their convenience during their leisure time.
- **Bias:** As there is no personal contact between researcher and respondents, there is unlikely the case for personal bias based on first impressions to alter the response of the survey. This is one of the advantages of mail survey because if the respondents do not like the approach of the interviewer, the survey results may be unfavourably affected.
- **Sampling-internal link:** As there is no direct personal contact between researcher and respondents, there is a high possibility of covering a wide range of geographical area for survey.

Disadvantages: However, the mail survey suffers from many disadvantages as given below.

- **Very low response rate:** Among all the survey methods, the response rate in the mail survey is the least. The response rate is expected to be 20% to 30%.
- **Ability of respondent to answer survey:** In the mail survey, we assume that physical ability, literacy level and language ability of the respondents are adequate to participate in the survey. As most of the surveys pull the respondents from a random sampling, it is impossible to control such variables. Many target respondents may have different primary language than that used in the survey. Similarly, many of the target respondents may be illiterate or with low-reading level and thus may fail to respond to the questions correctly. People with physical inability such as dyslexia, visual impairment or old age problems may not have the competence to complete the survey.

Telephonic survey

This is an alternative form of interview to the personal, face-to-face interview, where the interviewer collects the relevant information from the target respondents through telephonic conversation. Following points should be helpful to locate the respondents and make them agree to participate for the telephonic survey.

Locate the respondent:

- The repeat calls could be necessary if the respondents work in organizations and the only channel to reach them is through their secretaries.
- As the researcher may not know the name and designation of the respondents, there is every possibility of interviewing the wrong person.

■ In a telephonic survey, the interviewer collects the relevant information from the target respondents through telephonic conversation.

- The researcher can intimate the respondents in advance informing them about the telephonic interview on the subject matter.

Making them agree to take part:

- The purpose of the call should be stated clearly to the respondents just like the introductory letter of a postal questionnaire.
- Generally, respondents listen to the introduction before they decide to participate or refuse.
- The researcher should motivate the respondents in the right way if they raise objections about why they could not participate in the survey.

Advantages: Following are the list of advantages of telephonic survey.

- It is more flexible in comparison to mailing method.
- It is faster than other methods of survey. It is a quick way of obtaining the information.
- is relatively cheaper.
- It can cover reasonably large number of people or organizations with wide geographic coverage.
- High-response rate—keep going till the required number is reached.
- Recall is easy; call-backs are simple and economical.
- Interviewer can explain requirements more easily.
- Replies can be recorded without causing embarrassment to respondents.

Disadvantages: The disadvantages of telephonic survey are given below.

- This kind of survey is often connected with selling.
- The questionnaire must be short and questions must be simple and straightforward, otherwise respondents may refuse to answer them.
- Surveys are restricted to respondents who have telephone facilities.
- Repeat calls are inevitable—average 2.5 calls to get to someone.
- Time is wasted.
- Respondent has little time to think to respond to each question being asked over the phone.
- It is not possible to use visual aids.
- Too many questions or disturbance in the telephonic connection may cause irritation to the respondents.
- Not suitable for intensive surveys where comprehensive answers are required to various questions.

Internet survey

With the growth of the Internet and the expanded use of electronic mail for all purposes, the electronic survey is becoming one of the most widely used survey method these days. The electronic surveys can be done in many ways: (i) The survey forms can be distributed as electronic mail messages through attachment to potential respondents, (ii) the survey form can be posted as World Wide Web forms on the Internet, and (iii) the survey form can be distributed via publicly available computers in high-traffic areas such as libraries and shopping malls.

In the Internet survey, the questionnaire can be distributed as mail messages through attachments or posted as World Wide Web forms or distributed via publically available computers in high-traffic areas.

Advantages: The advantages of Internet survey are as follows.

- **Cost saving:** It is comparatively very less expensive to send questionnaires online than to pay for postage in a mail survey or for interviewers in a personal interview.
- **Ease of editing/analysis:** It is easier to make changes to the questionnaire and to copy and sort data.
- **Faster transmission time:** Unlike mail survey where it may take few days to deliver the questionnaire, the internet survey just takes few seconds to deliver the questionnaire to the respondents.
- **Easy use of pre-letters:** The researcher can estimate the response rate through the invitations. The active and interested participants may respond to the invitation in a very short time period, which may give a rough estimate of participation in the survey. You may send invitations and receive responses in a very short time and thus receive participation level estimates.
- **Higher response rate:** Research shows that response rates on private networks are higher with electronic surveys than with paper surveys or interviews.
- **More candid responses:** As there is no personal contact between researcher and respondents, the responses are unlikely to get biased by interviewer. Responses are expected to be more genuine with electronic survey than with a personal interview. Research shows that respondents may answer more honestly with electronic surveys than with paper surveys or interviews.
- **Potentially quicker response time with wider magnitude of coverage:** The electronic survey has the most widely geographic coverage of participants. The location of participant does not matter. Further, due to the speed of online networks, participants from any part of the world can answer in minutes or hours.

Disadvantages: However, Internet survey is not free from drawbacks. It has the following weaknesses.

- **Sample demographic limitations:** There are many potential participants who do not have access to computer and online network. So, electronic survey is limited to only those who have access to the Internet.
- **Lower levels of confidentiality:** It is difficult to guarantee anonymity and confidentiality because of the open nature of most of the online networks.
- **Layout and presentation issues:** Constructing the format of a computer questionnaire could be a difficult task for most of the researchers for few times due to lack of experience.
- **Additional orientation/instructions:** More detailed instruction and orientation to the computer online systems may be required for respondents to complete the survey.
- **Response rate:** It is observed that the response rate in an electronic survey is higher during first few days. Thereafter the response rate declines rapidly.

6.1.3 Qualitative Techniques

Sometimes, the research objective calls for more indirect methods of questioning, either because normal quantitative surveys are inadequate or

inappropriate. In such cases, qualitative methods, which probe the mind of respondent, might be useful. The major requirement for qualitative techniques is that we need a behavioural specialist such as a psychologist or sociologist to analyze the findings. The sample size in qualitative techniques is usually small, and analysis and interpretation is not as easy as it is in the quantitative studies. If done by non-expert, qualitative research can be completely misleading. Some of the important qualitative tools are discussed in detail below.

In-depth interview

In-depth interviewing is a qualitative research technique that involves conducting intensive individual interviews with a small number of respondents to explore their perspectives on a particular idea, programme or situation. For example, we might ask participants, staff and others associated with a programme about their experiences and expectations related to the programme, the thoughts they have concerning programme operations, processes and outcomes, and about any changes they perceive in themselves as a result of their involvement in the programme. In-depth interviews are useful when detailed information about a person's thoughts and behaviours is needed or new issues are to be explored in depth. Interviews are often used to provide context to other data (such as outcome data), offering a complete picture of what happened in the programme and why. For example, you may have measured an increase in youth visits to a clinic, and through in-depth interviews you find out that a youth noted that she went to the clinic because she saw a new sign outside of the clinic advertising youth hours. You might also interview a clinic staff member to find out their perspective on the clinic's "youth friendliness".

■ In-depth interviewing is a qualitative research technique that involves conducting intensive individual interviews with a small number of respondents to explore their perspectives on a particular idea, programme or situation.

In-depth interviews should be used in place of focus groups if the potential participants are not to be included or are not comfortable talking openly in a group, or when one wants to distinguish an individual's (as opposed to group) opinions about the programme. They are often used to refine questions for future surveys of a particular group.

The primary advantage of in-depth interviews is that they provide much more detailed information than what is available through other data collection methods such as surveys. They also may provide a more relaxed atmosphere in which to collect information—people may feel more comfortable having a conversation with you about their programme as opposed to filling out a survey. However, there are a few limitations and pitfalls, each of which is described below.

- The responses to interview community members and programme participants could also be biased due to their stake in the programme or for a number of other reasons. Every effort should be made to design a data collection effort, create instruments and conduct interviews to allow for minimal bias.
- The interviews can be a time-intensive evaluation activity as it consumes long time to conduct interviews, transcribe them and analyze the results.

- The interviewer must be appropriately trained in interviewing techniques. To provide the most detailed and rich data from an interviewee, the interviewer must make that person comfortable and appear interested in what they are saying. They must also be sure to use effective interview techniques, such as avoiding yes/no and leading questions, using appropriate body language and keeping their personal opinions in check.
- When in-depth interviews are conducted, generalizations about the results are usually not able to be made because small samples are chosen and random sampling methods are not used. In-depth interviews however, provide valuable information for programmes, particularly when supplementing other methods of data collection. It should be noted that the general rule on sample size for interviews is that when the same stories, themes, issues and topics are emerging from the interviewees, then a sufficient sample size has been reached.

■ The focus group is a carefully planned and moderated discussion among individuals having common characteristics to obtain the meaningful information on the area of interest in a non-threatening environment.

Focus group discussion

A focus group is a group of interacting individuals having some common interests or characteristics, brought together by a researcher, who uses the group and its interaction as a way to gain information about a specific or focused issue. The focus group is a carefully planned and moderated discussion to obtain the meaningful information on the area of interest in a non-threatening environment. The focus group discussion is an unstructured method of data collection where the respondents express their views freely. It is mostly used for explorative studies rather than any conclusive studies. Groups are comprised of respondents who share similar concerns and responsibilities but have minimal contact with each other in their daily lives. As groups differ in their composition and dynamics, multiple groups should be organized to obtain information from a different perspective on a given topic. Groups typically contain approximately 6–12 people: large enough to provide for a range of views but small enough for everyone to contribute.

■ Projective techniques are used by psychologists for projections of respondents in inferring underlying motives, urges or intentions.

Projective techniques

Projective techniques are used by psychologists for projections of respondents in inferring underlying motives, urges or intentions which are such that the respondents resist revealing them or are unable to figure it out themselves. The respondent in supplying information tends unconsciously to project his/her own attitudes or feelings on the subject under study. Projective techniques play an important role in motivational studies or in attitude surveys. In the following paragraphs, we shall discuss some of the important projective techniques.

- 1 **Word association:** In this technique, we ask the respondents to associate brands with one word that they think of when they think of the brand. It could also be a person, a celebrity or an animal, depending on the interviewer's or the analyst's view point. Interpretation of such association is best left to a psychologist, or a researcher with a psychoanalytical background and experience.

- 2 **Sentence completion:** Under this, respondent is given an incomplete sentence (such as, "People who go to shopping mall tend to be....") to complete to find an association of shopping mall with certain personality characteristics. Similarly several sentences might be put to the respondents on the same subject. Analysis of replies from the same respondent reveals his attitudes, and the combination of these attitudes of all the respondents in the sample is then taken to reflect the views of the population. This technique permits the testing not only of words (as in case of word association tests), but of ideas as well and thus, helps in developing hypotheses and in the construction of a questionnaire.
- 3 **Verbal projection:** In this technique, the respondents are asked to comment on or to explain what other people do. For example, why do people smoke? Answers may reveal the respondent's own motivation towards smoking.

6.2 Collection of Secondary Data

Secondary data is indispensable for most organizational research. Secondary data refers to information that have been already gathered by someone (individual or agencies) and readily available to the researcher. The secondary data can be internal or external to the organization and it can be accessed through the Internet or perusal of recorded or published information.

Generally, business research should be undertaken after a prior search of secondary sources. The secondary sources of information are important for any business research due to the following reasons.

- Secondary data may be available which is entirely appropriate and wholly adequate to draw conclusions and answer the question or solve the problem. Sometimes primary data collection simply is not necessary.
- It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget, a thorough examination of secondary sources can yield a great deal of more information than can be gathered through a primary data collection exercise.
- The time involved in searching secondary sources is much less than that needed to complete primary data collection.
- Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has undertaken a large-scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.
- It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses. The assembly and analysis of secondary data almost invariably improves the researcher's understanding of the marketing problem, the various lines of inquiry that could or should be followed and the alternative courses of action which might be pursued.

- Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken. For instance, government statistics on a country's agriculture will help decide how to stratify a sample and, once sample estimates have been calculated, these can be used to project those estimates to the population.

There are several sources of secondary published data and available in (a) various publications of the central, state and local governments such as publications of economic indicators, census data and statistical abstracts; (b) various publications of foreign governments or international bodies and their subsidiary organizations; (c) books and periodicals; (d) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (e) reports generated by research scholars of academic institutions in various fields; and (f) public records and statistics, historical documents and other sources of published information. The secondary data also could be from unpublished sources such as unpublished biographies and autographies, unpublished research thesis, working research papers of scholars, etc.

Researchers must be very careful in using secondary data because it is just possible that the available data may not be suitable or may be inadequate in the context of the problem under investigation. By way of caution, the researcher, before using secondary data, must see that they possess following characteristics.

- 1 **Reliability of data:** The reliability can be tested by investigating such things about the said data: (a) Who collected the data? (b) What were the sources of data? (c) Was the data collected by using appropriate methods? (d) At what time period data was collected? (e) What level of accuracy was desired and how far it was achieved?
- 2 **Suitability of data:** The data that are suitable for one enquiry may not necessarily be found suitable for another enquiry. The researcher must carefully scrutinize the definition of various terms and units of collection used in the study before identification of relevant data from the published sources.
- 3 **Adequacy of data:** If the level of accuracy in data is found inadequate for the purpose of present enquiry, data will be considered as inadequate and should not be used by a researcher. The data will also be considered as inadequate, if they are related to an area which may be either narrower or wider than the area of the present study.

The available data from secondary sources should be used by researchers when he/she finds data to be reliable, suitable and adequate. But we should not blindly discard the use of such data if it is readily available from the secondary sources as it will not be economical to spend time and energy in field survey for collecting information.

6.3**Selection of Appropriate Methods of Data Collection**

We have discussed various methods of data collection. Each method has its own advantages and disadvantages. The researcher must judiciously select the method(s) for his/her own study keeping in view the following factors.

- 1 **Nature, scope and object of enquiry:** This is the most important factor affecting the choice of a particular method of data collection. The method selected should be such that it suits the type of investigation that is to be conducted by researchers. It is also important to decide whether the required data is available from the secondary sources or need to be collected through the primary sources.
- 2 **Availability of funds:** The choice of data collection method depends very much on the availability of funds for the research. When the fund is limited, the researcher may be forced to choose relatively cheaper method which may not be as effective as some other method which requires high budget.
- 3 **Time factor:** The time at the disposal of the researcher do affects the choice of data collection method. Some methods are relatively more time consuming than other methods. So the time factor should be taken into consideration by the researchers when they plan for any particular method of data collection.
- 4 **Precision required:** The researcher should give the consideration on the level of precision required while selecting the method of data collection..

One should always keep in mind that each method of data collection has its own advantages and disadvantages and none is superior in all situations. For instance, telephonic interview method may be considered appropriate if funds are restricted, time is also restricted and the data is to be collected in respect of very few items with or without a certain degree of precision. In case funds permit and more information is desired, personal interview could be relatively better method. In case time is ample, funds are limited and much information is to be gathered with no precision, then mail survey can be regarded more reasonable. When a wide geographical area is to be covered, the use of mail survey supplemented by personal interview will yield more reliable results per RM spent than either method alone. The secondary data can be used when the researcher finds them reliable, adequate and appropriate for the problem in hand. While studying motivating influences in the market researches or studying people's attitude in psychological and social surveys, we can resort to the use of one or more of the projective techniques.

Thus, the most desirable approach with regard to the selection of the method depends on the nature of the particular problem and on time and resources available with the desired degree of accuracy. But over and above all this, much depends upon the ability and experience of the researchers.

Each method of data collection has its own advantages and disadvantages. The selection of a particular method depends on many factors such as nature and scope of the study, availability of funds, and time and the level of precision required.

The researcher should be concerned on whether one's procedures of collecting information are likely to cause any physical or emotional harm to the participants.

6.4 Ethical Issues

Several ethical issues should be addressed while collecting data. We should be concerned on whether one's procedures of collecting information are likely to cause any physical or emotional harm. These harms may be caused by:

- violating participants' right to privacy by posing sensitive questions or by gaining access to records which may contain personal data;
- observing the behaviour of participants without their being aware (concealed observation should therefore always be crosschecked or discussed with other researchers with respect to ethical admissibility);
- allowing personal information to be made public which participants would want to be kept private; and
- failing to observe/respect certain cultural values, traditions or taboos valued by the participants.

Several methods for dealing with these issues may be recommended:

- obtaining the respondent's consent before the study or the interview begins;
- not exploring sensitive issues before a good relationship has been established with the participant;
- ensuring the confidentiality of the data obtained; and
- learning enough about the culture of participants to ensure it is respected during the data collection process.

If sensitive questions are asked, for example, about family planning or sexual practices, or about opinions of patients on the health services provided, it may be advisable to omit names and addresses from the questionnaires.

Chapter Summary

Data collection techniques allow us to systematically collect information about objects of study (people, objects, phenomena) and about the settings in which they occur. Data can be collected through primary or secondary research. Secondary research involves any information from the published sources which has not been specifically collected for the current research problem. This includes the published sources of data such as an electronic database, periodicals, company's annual reports, government report, etc. On the other hand, primary research involves collecting information specifically for the study on hand from the actual sources such as consumers, users/non-users or other entities involved in research.

Broadly speaking, there are basically two methods of data collection—qualitative and quan-

titative. Qualitative research explores attitudes, behaviour and experiences through such methods as interviews or focus groups. It attempts to get an in-depth opinion from participants. Quantitative research generates statistics through the use of large-scale survey research; using methods such as questionnaires or structured interviews. Some of the important and most widely used methods of data collection include observation, survey method, in-depth interviews, focus group discussion and projective techniques.

The observation method is the most commonly used method especially in studies relating to behavioural sciences. The observation is a technique that involves systematically selecting, watching and recording behaviour and characteristics of

living beings, objects or phenomena. The survey method involves the collection of information from representative target respondents using a pre-designed questionnaire. Unlike observation, it is structured method of data collection in which we extract exactly the same information from all the target population. Survey can be done in many ways such as through personal interview, mail survey, telephonic survey and internet survey. In-depth interviewing is a qualitative research technique that involves conducting intensive individual interviews with a small number of respondents to explore their perspectives on a particular idea, programme or situation. The focus group discussion is another qualitative technique where a group of interacting individuals having some common interest or characteristics are brought together by a researcher, who uses the group and its interaction as a way to gain information

about a specific or focused issue. Finally, projective techniques are used by psychologists to use projections of respondents for inferring about underlying motives, urges or intentions which are such that the respondents resist revealing them or are unable to figure out themselves.

Each method of data collection has its own advantages and disadvantages and none is superior in all situations. The choice of a particular method depends on many factors such as nature and scope of investigation, availability of funds and time, and precision required. Whatever may be the choice of method, the ethical issues should be addressed while collecting data. We should be concerned on whether one's procedures of collecting information are likely to cause any physical or emotional harm.

Discussion Questions

1. What is the difference between qualitative and quantitative methods of data collection?
2. What is the difference between the observation and survey method?
3. Explain the advantages and disadvantages of various survey methods.
4. Describe some of the major projective techniques and evaluate their significance as a tool of business research.
5. Write short notes on the following methods of data collection:
 - (a) In-depth interview.
 - (b) Focus group discussion.
6. Discuss the importance of secondary data. What are some cautions a researcher should take while using secondary data?
7. Discuss the factors that affect the choice of data collection method.



CHAPTER

7

Measurement in Management and Scaling Techniques

Learning Outcomes

After reading this chapter, student should be able to:

- Understand the four basic measurement techniques in business research
- Learn different measurement scales under comparative and non-comparative scaling techniques. Select the correct measurement scales for different types of statements in the questionnaire and take a number of practical decisions into account while developing the scale for questions
- Test the measurement instruments for its degree of stability, consistency and reliability
- Test the measurement instrument for content validity, face validity, construct validity and criterion validity

7.0 Introduction

We must correctly measure the concepts we are examining. Otherwise our interpretations and conclusions will be misleading or inaccurate. In our day to day life, we measure products, performance, quality and so on. For example, *exam* measures our achievement, *quarterly or annual reviews* measures our progress, *metre* measures the size, *litre* measures the quantity of petrol and *kilogram (kg)* measures the weight. Sometimes, subconsciously we measure something. We buy an ice cream cone and say it tastes good. We are measuring the quality of the ice cream. We

say about an employee that he is lazy, irresponsible or uncooperative. *We are measuring the attitude of the employee.*

Measurement is fundamental to business research. Without appropriate measurement, it is difficult, if not impossible, to comment on business behaviour or business phenomenon.

- HR managers measure employees' performance, motivation, turnover and similar indices.
- Accountants measure profits and losses, assets and liabilities, depreciation and so on.
- Marketing managers measure the service quality, brand image, brand preference and so on.

The more effectively managers measure these business aspects, the better their decision.

A scale of measurement allows the investigator to make comparisons of amounts and changes in the variable being measured. Measurement consists of two basic processes called conceptualization and operationalization. First, variables are defined by conceptual definitions (constructs) that explain the concept the variable is attempting to capture. Secondly, variables are defined by operational definitions, that is, definitions of how variables will be measured. It is followed by an advanced process called determining the levels of measurement, and measuring reliability and validity of an instrument which is the focus of the current chapter.

7.1 Basic Measurement Techniques

Level of measurement is important while measuring the variables. The higher the level of measurement of a variable, the more powerful is the statistical techniques that can be used to analyze it. There are four basic types of measurement in business research. They are nominal, ordinal, interval and ratio.

Nominal scale

Nominal measurement consists of assigning items to groups or categories. It has following features:

- It is used to indicate categories.
- Numbers are only used as labels.
- It has no numerical significance.
- It does not represent any order or distance.

Religious preference, race and sex are all examples of nominal scales. For example, variable gender and race can be coded as follows:

Gender	Code
Male	1
Female	2

The four basic measurements of scale are:

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal scale consists of categorizing items into groups.

Race	Code
Malay	1
Chinese	2
Indian	3
Others	4

In the above examples, the numerical values do not signify anything. It is only an indication of category. Thus, these numbers can be interchanged. For instance, instead of putting the value 1 for male and 2 for female, we can assign the value 1 for female and 2 for male. Similarly we can code *Malay* as 4, *Chinese* as 3, *Indian* as 2, and *Others* as 1 or as you desire. The only thing is that one needs to remember what numerical value stands for which category.

Nominally scaled variables cannot be used to perform many statistical computations such as mean and standard deviation, because such statistics do not have any meaning when used with nominal scale variables. With nominally scaled variables, the analysis is confined to frequency and cross-tabulation. The chi-square test can be performed on a cross-tabulation of nominal scale data.

Ordinal scale

The main characteristic of the ordinal scale is that the categories have a logical or ordered relationship to each other. These types of scale permit the measurement of degrees of difference, but not the specific amount of difference. This scale is very common in marketing, satisfaction and attitudinal research.

- Ordinal scale implies "ranking" on the basis of preference.
- It does not say anything about the "distance".
- Ranks are not interchangeable as nominal scale labels are.

Example 7.1: On what basis do you select an electronic product? On the basis of preference, please rank the following attributes.

Attributes	Ranking	Ranking (R1)
Popularity		5
Brand Name		1
Image		3
Current trends		2
Availability		4

In the above example, the rating on the above attributes of a respondent (R1) is given in the third column. As per the score given by R1, the brand name is the most important attribute in making his/her decision to select an electronic product. It is followed by current trends, image, availability and popularity. Unlike, nominal scale, here the numbers are assigned by the respondents and thus cannot be interchanged as it will change the preference of the respondent. The number only indicates the preference (one attribute is preferred to another). For R1, "brand name" is preferred to any other attributes. The attribute "image" is preferred to availability and popularity of the product.

and so on. However, these numbers do not measure the degree of differences in the preferences. Looking at these numbers, we cannot say "brand image" is five times more important than "popularity" for R1 in making a decision to buy the product.

In addition to frequency tabulation and cross-tabulations, the other statistics that can be used with the ordinal scale are median, various percentiles such as quartile and rank correlation. The arithmetic mean should not be used as the average ranking does not make any sense here.

The ordinal scale implies ranking based on preference.

Interval scale

The interval scale implies rating on a particular scale.

The interval scale is also known as *rating scale* and variables (attributes) are measured on different scales such as scale of 1 to 5 or 1 to 7 or 1 to 10. In this scale, it is assumed to have equidistant points between each of the scale elements. This means that we can interpret differences in the distance along the scale. In contrast to an ordinal scale where we can only know about differences in order, not differences in the degree of order.

Example 7.2: Rate the following brands of detergent soap in their ability to clean clothes on a scale of 1 to 7 (1 = very low ability, 7 = very high ability). Please circle the number which appropriately reflects your answer.

Brand									
Ariel	1	2	3	4	5	6	7		
Breeze	1	2	3	4	5	6	7		
Tide	1	2	3	4	5	6	7		

In the above example, the rating by a respondent R1 is shown in circle. The most important brand of detergent in terms of its ability to clean is Ariel, followed by Tide and Breeze. The numbers here not only talk about the preference, but also measure the distance. For example, Ariel is 2 times better than Breeze in terms of its ability to clean cloths.

The interval data can be used to calculate mean, standard deviation, correlation coefficient, regression, analysis of variance, factor analysis, and a whole range of advanced multivariate and modelling techniques.

Ratio scale

The ratio scale possesses all the properties of nominal, ordinal and interval scale, and in addition, an absolute zero point.

A ratio scale possesses all the properties of the nominal, ordinal and interval scale, and in addition, an absolute meaningful zero point. For example, age is an example of ratio scale. If we ask respondents their ages, the difference between any two years would always be the same, and 'zero' signifies the absence of age or birth. Hence, an 80-year-old person is indeed twice as old as a 40-year old one. Physical measurements such as height, weight and length are typically ratio variables. Similarly, years of participation, sales figures, quantities purchased and market share are all expressed on a ratio scale.

Table 7.1 provides the differences among the four basic scales discussed above with the help of an example.

Nominal Numbers assigned to runners	Runner No. 12	Runner No. 9	Runner No. 18
Ordinal (Rank order of winner)	3rd Place	2nd Place	1st Place
Interval (Performance rating on 1-10 scale)	7	8	9
Ratio (Time to finish in seconds)	16.2	14.6	13.5

Table 7.1

Differences among basic measurement scales: An example

Let us take an example of "100 meter race" in a tournament where three runners are participating from three different states of Malaysia. Each runner is assigned a number (displayed in uniform) to differentiate from each other. The number displayed in the uniform (row 1) to identify runners is an example of nominal scale. Once the race is over, the winner is declared along with the declaration of first runner up and second runner up based on the criteria that who reaches the destination first, second and last. The rank order of runners (row 2) such as "second runner up as 3", "first runner up as 2" and the "winner as 1" is an example of ordinal scale. During the running tournament, judge is asked to rate each runner on the scale of 1-10 based on certain criteria. The rating given by the judge (row 3) is an example of interval scale. The time spent by each runner (row 4) in completing the race is an example of ratio scale.

The ratio scale is the top level of measurement and it can be used for all the statistics as in the case of interval scale.

7.2 Advanced Measurement Techniques

The various types of scaling techniques used in research can be classified into two categories: (i) comparative scales and (ii) non-comparative scales. In comparative rating scale, the respondent evaluates (assign a value) to a particular item/brand/product in comparison to other items/brands/products. The evaluation cannot be possible without comparison. On the other hand, in non-comparative scaling respondents evaluate each item/brand/product independently without comparing with any other items/brands/products.

Spot the difference

Q1. Please rank the following brands of toothpaste based on your preference.

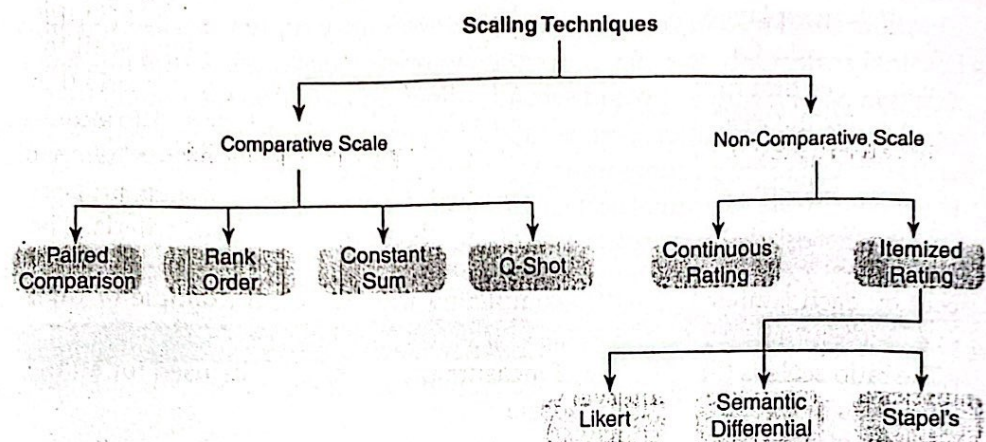
Brand	Rank
Colgate	
Pepsodent	
Darlie	
Fresh & White	
Sensodyne	

Q2 Please rate the following brands of toothpaste on the scale of 1 to 5 as given below:

Brand	Very much dislike	Dislike	Neither like nor dislike	Like	Very much like
Colgate	1	2	3	4	5
Pepsodent	1	2	3	4	5
Darile	1	2	3	4	5
Fresh & White	1	2	3	4	5
Sensodyne	1	2	3	4	5

In order to evaluate the brand "Colgate" in Q1, the respondents need to compare this brand with the rest of four brands. However, the respondents can evaluate the brand "Colgate" in Q2 independently without comparing it with the rest of brands.

Figure 7.1
Different types of scaling techniques



7.2.1 Comparative Scales

The comparative scales can be further divided into the following four types of scaling techniques: (i) paired comparison scale, (ii) rank order scale, (iii) constant sum scale and (iv) Q-sort scale.

Paired comparison scale

This is a comparative scaling technique in which a respondent is presented with two objects at a time and asked to select one object (rate between two objects at a time) based on some criterion. The data obtained is ordinal in nature. A respondent has to make $[n(n-1)/2]$ paired comparison for n number of items. For example, if there are five brands of mobile handsets (Nokia, LG, Motorola, Samsung, Sony) to be evaluated, a respondent has to make $[5(5-1)/2] = 10$ paired comparisons as given below:

- 1 Nokia – LG
- 2 Nokia – Motorola
- 3 Nokia – Samsung
- 4 Nokia – Sony

The types of comparative scaling techniques that are frequently used are paired comparison, rank order, constant sum and Q-sort.

- 5 LG – Motorola
- 6 LG – Samsung
- 7 LG – Sony
- 8 Motorola – Samsung
- 9 Motorola – Sony
- 10 Samsung – Sony

The following is the data recording format using the paired comparison:

	Nokia	LG	Samsung	Motorola	Sony
Nokia	X	0	0	0	1
LG		X			
Samsung	1	0	X	0	1
Motorola		0		X	
Sony	0	0	0	0	X
No. of times preferred	3	0	2	1	4
Rank	2	5	3	4	1

"1" in a particular box means that the brand in that column is preferred over the brand in the corresponding row. "0" in a particular box means that the brand in that row is preferred over the brand in the corresponding column. In the above recording, Nokia is preferred over LG, Samsung and Motorola. However, Nokia is not preferred over Sony. The sum of column provides the number of times a particular brand is preferred over other brands.

In the above data input format, the response of a respondent R1 is recorded, which indicates that the brand "Sony" is preferred most, i.e., 4 times, followed by the brand "Nokia" with 3 times, and brand "Samsung" with 2 times. The brand "Motorola" is preferred only once and the brand "LG" is never preferred by the respondent R1.

Advantages

- Some special techniques such as multidimensional scaling require the data to be collected based on pair comparison.

Disadvantages

- Data obtained is ordinal.
- If the number of objects (n) is large, there is a high risk of ill considered answers or refusal to answer.

Rank order

A respondent is presented with several objects simultaneously and ask to order or rank them according to some criterion.

Example 7.3: Rank the various brands of toothpaste in order of preference. Begin by picking out the one brand that you like the most and assign it a number 1. Then find the second most preferred brand and assign it a number 2. Continue this procedure until you have ranked all brands of toothpaste in order of preference.

Brand	Rank
Colgate	—
Close Up	—
Pepsodent	—
Plus White	—
Darlie	—
Fresh & White	—

Advantages

- It takes less time as compared to "paired comparison scaling". If there are "n" stimulus objects, only (n-1) scaling decisions need to be made.
- Rank order scaling is commonly used to measure preference of the brand as well as attributes.
- Rank order data is obtained in conjoint analysis.

Disadvantages

- Rank order scaling results in ordinal data.

Constant sum scaling

A respondent allocates a constant sum of units (usually points) among a set of stimulus objects with respect to some criterion. If an attribute is unimportant, the respondent may assign it zero point. If an attribute is twice as important as some other attribute, it receives twice as many points.

Example 7.4: Below are eight attributes of toilet shop. Please allocate 100 points among the attributes so that your allocation reflects the relative importance you attach to each attribute.

Attribute	Response (points)
Mildness	12
Lather	11
Shrinkage	7
Price	45
Fragrance	13
Packaging	4
Moisturizing	8
Sum	100

Advantages

- It allows fine discrimination among stimulus objects without requiring too much time.
- The constant sum scaling not only measures the preference but also the degree to which a particular attribute is more important than others.

Disadvantages

Respondents may allocate more or fewer units than those specified.

Q-Sort scale

It uses a rank order procedure and the objects are sorted into piles based on similarity with respect to some criteria. The number of objects to be sorted

Thus, the Q-sort technique is an attempt to classify subjects in terms of their similarity to attribute under study.

Non-comparative scales are further divided into following four types: (i) continuous rating scale, (ii) Likert scale, (iii) semantic differential scale and (iv) Stapel's scale.

This is the oldest and most widely used method for performance appraisal. A respondent is asked to rate the objects by placing a mark at the appropriate position on a line that runs from one extreme criterion variable to other. A respondent is not restricted to selecting from marks previously set by the researcher. It is also known as *graphic rating scale*.

0 1 2 3 4 5 6 7 8 9 10
(Worst) (Best)

Itemized rating scale

Each statement is assigned a numerical score ranging from either -2 to +2 or 1 to 5. Total score of each respondent is calculated by summing across the item.

1 = Strongly disagree
2 = Disagree
3 = Neither agree nor disagree
4 = Agree
5 = Strongly agree

Scanned with CamScanner

- Each item has only one word/phrase indicating the dimension it represents.
- Each item has ten response categories.
- Each item has an even number of categories.
- The response categories have numerical labels but no verbal labels.

Example 7.8: Measuring the attitude of flight passengers

+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
Friendly Cabin Crew	Comfortable Interior	Accurate Timing
-2	-2	-2
-3	-3	-3
-4	-4	-4
-5	-5	-5

7.2.3 Practical Consideration

While developing the scale for questions, a number of practical decisions should be taken into account. They are discussed below.

Number of scale categories

One of the important decisions that one has to consider is whether one should have 5 point scale or 7 point scale or 10 point scale. From a research design perspective, the larger the number of categories, the greater the precision of the measurement scale. But it may be more difficult to discriminate between levels and respondents may face greater difficulty in processing information if the number of categories is more. Thus the desire for a higher level of precision must be reasonably balanced with the demands placed on the respondents.

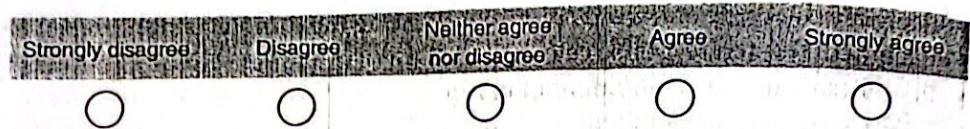
Number of items to measure concept

Concepts are measured using scales with multiple items known as multi-item scales. A multi-item scale consists of a number of closely related individual statements (items) whose responses are combined into a composite score. But the question is what should be the appropriate number of items to measure a particular concept? The general guideline is that statements must be closely related, represent only a single construct, and must completely represent the construct to be measured with the multi-item scale. Generally it is common to see five to seven items and even more to measure a single concept. A minimum of three items is must to achieve acceptable reliability.

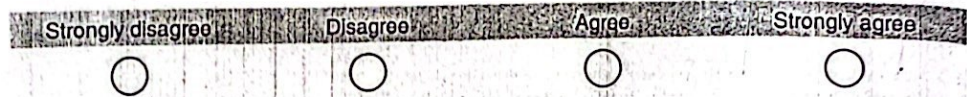
■ Minimum three items are required to measure a concept.

Odd or even number of categories

There is debate on whether one should use even or odd numbers of points on a rating scale. In an odd number of categories of a scale, the mid-point represents a neutral position. The scale below is an example of an odd one.



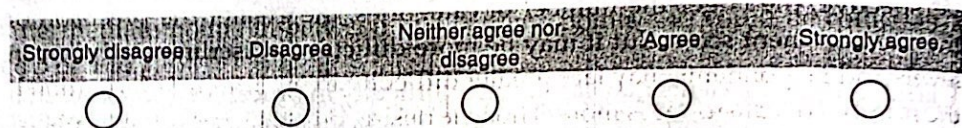
This type of scale is generally used when, based on the experience or judgement of the researchers, it is believed that a part of the sample are likely to feel neutral about the issue being examined. However, there are reasons a researcher might prefer an even scale over an odd one as given below.



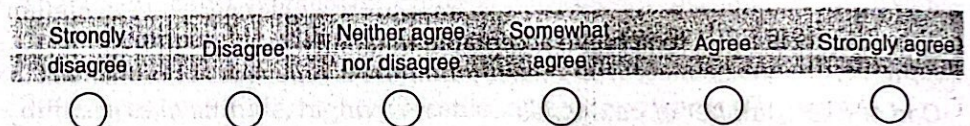
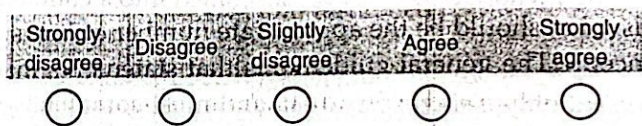
In the above scale, the respondents are forced to draw their stand about a particular point of view. This scale could be useful to implement if the question is on something a respondent cannot be undecided about or if an issue is such that respondents are biased towards the neutral option. By omitting the neutral option in the middle of the scale, the respondents cannot "cop-out" by choosing it. However, forcing respondents to choose either positive or negative options may cause some of them to skip the question or force them not to reveal the true response to the question

Balanced scale or unbalanced scale

Scales can be either balanced or unbalanced. The scale is known as balanced if the number of positive (favourable) options is equal to number of negative (unfavourable) options. The scale given below is an example of balanced scale.



On the other hand, the scale is known as unbalanced if the number of positive options is greater than the number of negative options or the number of positive options is less than the number of negative options. The scales given below are examples of unbalanced scale.



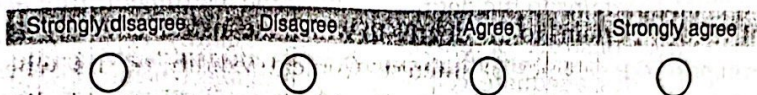
In the first example, the number of negative options is greater than the number of positive options, whereas in the second example, the number of negative options is less than the number of positive options.

Generally, rating scales should be balanced, with an equal number of favourable and unfavourable response choices. One cannot justify an unbalanced rating scale where there is no reason to believe that subjects are just as likely to be negative as positive. The only justification for using an unbalanced rating scale is in a situation where it is known *a priori* that virtually all respondents are leaning in one direction, e.g. brand loyal customers would be expected to be essentially favourable. If you know that one side of the scale will not really be used, you would then want the precision on the side of the scale that will be used. Therefore, an unbalanced rating scale might be called for.

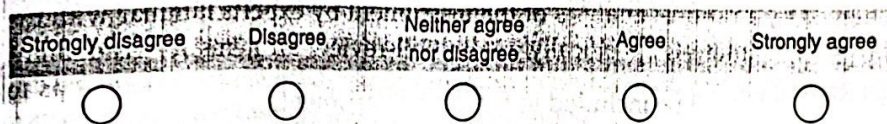
■ The unbalanced scale may be used in a situation when it is known *a priori* that virtually respondents are leaning in one direction.

Forced choice or non-forced choice

The discussion on even-point scales leads us into a discussion of forced and unforced choice questions. The four-point scale below is an example of forced choice.



In this case, the respondents are forced to either agree or disagree. However, the five-point scale below is an example of unforced choice question.



In this case, the respondents are not forced to choose either favourable or unfavourable response. If they want, they are allowed to be neutral with their response on the question. However, one can use the even-point scale with an unforced choice by placing an option for "Don't Know" or "Not Sure" after the "strongly disagree" option and then code it with a "DK" or an "X", rather than a "5".

A forced-choice rating scale will bias results by eliminating the undecideds and/or those with no opinion. Some researchers will purposely leave out the response choice of "undecided", "no opinion", "not sure" or "don't know". This approach may be reasonable when the researcher has good reason to believe that virtually all subjects have an opinion and you do not want them to "cop out" by indicating they are uncertain. What happens if many subjects are indeed undecided and we do not allow them the option of no opinion? Most will probably select a rating from the middle of the scale, e.g. "average" or "fair". This will cause two biases: (a) it will appear that more subjects

■ A forced-choice rating scale may bias results by eliminating the respondents who are truly neutral or have no opinion on the statement.

have opinions than actually do and (b) the mean and median will be shifted toward the middle of the scale. (The "undecided" category is not part of the scale.)

7.3 The Characteristics of Good Measurement

Sound measurement must meet the tests of validity, reliability and practicality. In fact, these are the three major considerations one should use in evaluating a measurement tool. Validity refers to the extent to which a test measures what we actually wish to measure. Reliability has to do with the accuracy and precision of a measurement procedure. Practicality is concerned with a wide range of factors of economy, convenience and interpretability.

7.3.1 Test of Validity

Validity is defined as the extent to which the instrument measures what it claims to measure. For example, a test that is used to screen applicants for a job is valid if its scores are directly related to future job performance.

■ The validity is the extent to which a test measures what we actually wish to measure.

Content validity

Content validity pertains to the degree to which the instrument fully assesses or measures the construct of interest. Content validity occurs when the experiment provides adequate coverage of the subject being studied. For example, say we are interested in evaluating the service quality of banks. We would want to ensure that our questions fully represent the domain of service quality of banks. The development of a content valid instrument is typically achieved by a rational analysis of the instrument by experts familiar with the construct of interest. Specifically, raters will review all items for readability, clarity and comprehensiveness, and come to some level of agreement as to which items should be included in the final instrument.

Face validity

The term face validity has a similar meaning. However, face validity generally refers to "non-expert" judgements of individuals completing the instrument and/or executives who must approve its use. Respondents may refuse to cooperate or may fail to treat seriously measurements that appear irrelevant to them. Managers may refuse to approve projects that utilize measurements lacking in face validity. Therefore, to the extent possible, researchers should strive for face validity.

Criterion validity

The criterion validity relates to our ability to predict some outcome or estimate the existence of some current condition. This form of validity reflects the success of measures used for some empirical estimating purpose. Any criterion must be judged based on four qualities: (i) relevance, (ii) freedom from bias, (iii) reliability and (iv) availability.

A criterion is *relevant* if it is defined in terms we judge to be the proper measure. For example, if you believe sales success is adequately measured

by Malaysian Ringgit Sales volume achieved per year, then it is a relevant criterion. If you believe success should include a high level of penetration of large accounts, then sales volume alone is not fully relevant.

Freedom from bias is attained when the criterion gives each subject an equal opportunity to score well. For example, the sales criterion would be biased if it did not show adjustments for differences in territory potential and competitive conditions.

A *reliability* criterion is stable and reproducible. An erratic criterion (usually monthly sales, which are highly variable from month to month) can hardly be considered a reliable standard to judge performance of salespersons.

Finally, the information specified by the criterion must be *available*. If it is not available, how much it will cost and how difficult it will be to secure? The amount of money and effort that should be spent on development of a criterion depends on the importance of the problem for which the test is used.

Criterion-related validity is expressed as the coefficient of correlation between test scores and some measure of future performance or between test scores and scores on another measure of known validity.

Construct validity

The construct validity is the most complex and abstract. In attempting to evaluate construct validity, we consider both the theory and the measuring instrument being used. A measure is said to process construct validity to the degree that it confirms to predicted correlations with other theoretical propositions. Construct validity is the degree to which scores on a test can be accounted for by the explanatory constructs of a sound theory. For determining construct validity, we associate a set of other propositions with the results received from using our measurement instrument. If measurement on our devised scale correlates in a predicted way with these other propositions, we can conclude that there is some construct validity.

Let a multi-item scale is developed to measure the tendency to purchase prestige brands. The theory suggests that this tendency is caused by three personality variables, i.e. (i) low self-focus, (ii) high need for status and (iii) high materialism. Further, we believe that "tendency to purchase prestige brands" is unrelated to "brand loyalty" and the "tendency to purchase new products". The evidence of construct validity would exist if our scale:

- 1 Correlates highly with other measures of prestige brand preference such as reported purchases and classifications by friends (*convergent validity*).
- 2 Has a low correlation with the unrelated constructs brand loyalty and tendency to purchase new products (*discriminant validity*).

7.3.2 Test of Reliability

The reliability of a measure indicates the extent to which it is without bias (error-free) and thus ensures consistent measurement across time and across the various items in the instrument. A measure is reliable to the degree that it supplies consistent results. Reliability is necessary contributor to validity but is not a sufficient condition for validity. The relationship between reliability and

■ The reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept.

■ The stability measures the degree at which the instrument provides consistent results with repeated measurements.

validity can be illustrated with a simple example of weighting instrument. If the weighting machine measures your weight correctly, then it is both reliable and valid. If it consistently overweighs you by 2 kg, then the scale is reliable but not valid. If the instrument measures erratically from time to time, then it is not reliable and therefore cannot be valid. The reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept and thus helps to assess the goodness of a measure.

Stability of measures

A measure is said to process stability if one can secure consistent results with repeated measurements of the same subject (respondent) with the same instrument. Two tests of stability are test-retest reliability and parallel-form reliability.

Test-retest reliability: The test-retest reliability estimates can be obtained by repeating the measurement using the same instrument under as nearly equivalent conditions as possible. The results of the two administrations are then compared and the degree of correspondence is determined. The greater the difference, the lower is the reliability. However, a number of practical and computational difficulties are involved in measuring test-retest reliability.

- (i) Some items can be measured only once. For example, it would not be possible to re-measure an individual's initial reaction to a new advertising slogan.
- (ii) The retaking of a measure may produce boredom, anger or attempts to remember the answers given in the initial measurement.
- (iii) Factors extraneous to the measuring process may cause shifts in the characteristic being measured. For example, a favourable experience with the brand during the period between the test and the retest might cause a shift in individual ratings of that brand.

Parallel-form reliability: The parallel-form reliability is used to assess the consistency of the results of two tests constructed in the same way from the same content domain. In this case we create a large set of questions that address the same construct and then randomly divide the questions into two sets and administer both instruments to the same sample of people. The results of the two administrations are then compared and the degree of correspondence is determined. The greater the difference, the lower is the reliability.

One major problem with this approach is that you have to be able to generate lots of items that reflect the same construct.

Internal consistency of measures

The internal consistency of measures is indicative of the homogeneity of items that measure the construct. The items should hang together as a set and be capable of independently measuring the same concept so that the subjects attach the same overall meaning to each of the items. This can be examined through the following two tests.

■ The internal consistency of measures is indicative of the homogeneity of items measuring the construct.

Inter-item consistency reliability: Inter-item consistency reliability also simply known as "internal consistency" measures the degree to which different items measuring the same construct attain consistent results. Scores on different items designed to measure the same construct should be highly correlated. The most popular test of inter-item consistency reliability is the Cronbach's coefficient of alpha.

Split-half reliability: Split-half reliability reflects the correlations between two halves of an instrument. The two halves can be created by splitting the items in several ways:

- (i) Odd and even numbered items;
- (ii) First and second halves, and
- (iii) Randomly.

If the results of the correlation are high, the instrument is said to have high reliability in an internal consistency sense. The high correlation tells us there is homogeneity among the items. The Spearman-Brown correlation formula is used to adjust the effect of test length and to estimate reliability of the whole test. However, it has certain limitations. Firstly, the estimates would vary depending on how the items in the measure are split into two halves. Secondly, the potential for incorrect inferences about high internal consistency exists when the test contains many items which inflate the correlation index.

7.3.3 Test of Practicality

From the operational point of view, the measuring instrument ought to be economical, convenient and interpretable. *Economy consideration* suggests that some trade-off is needed between ideal research project and that which the budget can afford. The length of the measuring instrument is an important area where economic pressure is quickly felt. More items give greater reliability, but at the cost of interview and observation time. So generally we limit the number of items for our study. Similarly the choice of method of data collection is dependent on economic factors. *Convenience* test suggests that measuring instrument should be easy to administer. One should give due attention to the proper layout of the instrument with clear instructions and coding. *Interpretability consideration* is more important when persons other than the designer of the instrument are to interpret the results. The measuring instrument in order to be interpretable must be supplemented by: (a) a detailed instruction for administering the test, (b) scoring keys, (c) evidence about the reliability and (d) guides for using the test and for interpreting the results.

■ From the operational point of view, the measuring instrument ought to be economical, convenient and interpretable.

Chapter Summary

Measurement is fundamental to business research. Without appropriate measurement, it is difficult, if not impossible, to comment on business behaviour or business phenomenon. There are four basic types of measurement in business research – nominal, ordinal, interval and ratio. Nominal measurement consists of assigning items to groups or categories. The ordinal scale implies ranking on the basis of preference, whereas the interval scale implies *rating* on user defined scale. The ratio scale possesses all the properties of nominal, ordinal and interval scale, and in addition, an absolute zero point.

The scaling techniques can be broadly classified into: (i) comparative scales, and (ii) non-comparative scales. In comparative scale, the respondent evaluates a particular item in comparison to other items, whereas in non-comparative scaling, respondent evaluates each item independently without comparing it with any other items. While developing the scale for questions, a number of practical considerations should be taken into account such as number of scale categories, number of items to measure concept, choice between odd and even number of

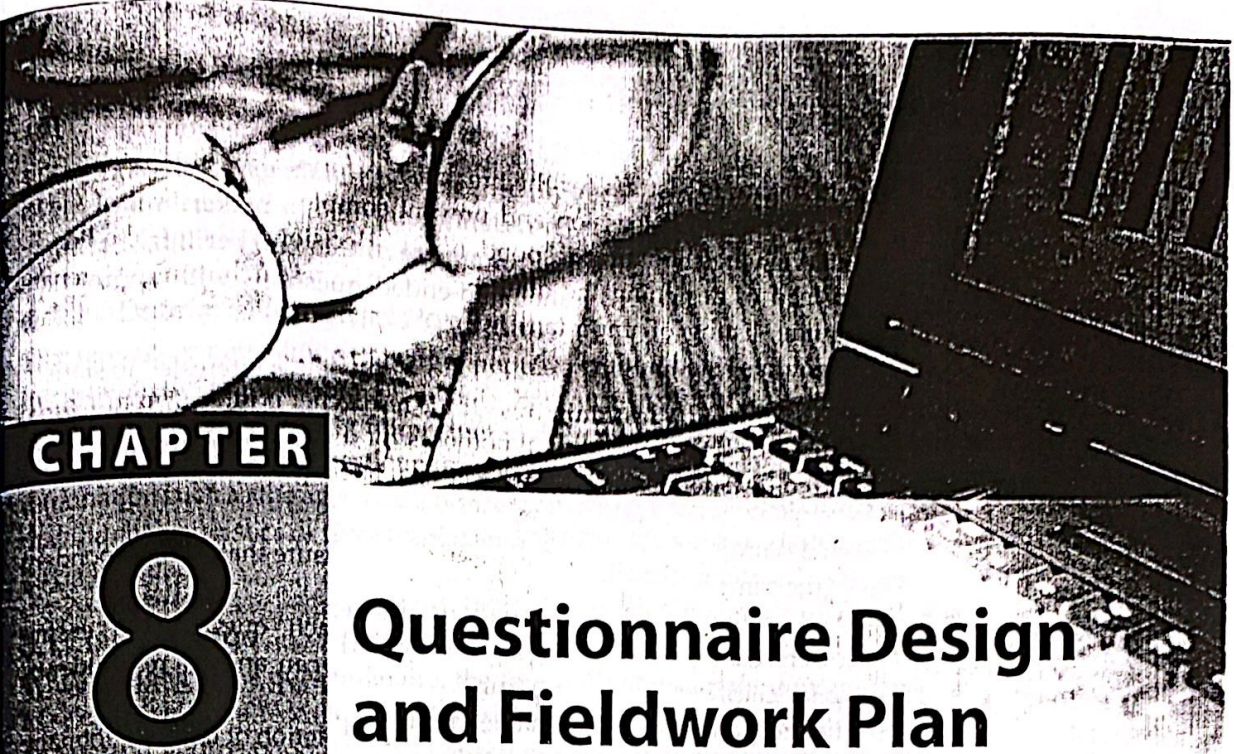
categories, balanced and unbalanced scale, forced and non-forced choice.

There are three major considerations one should use in evaluating a measurement tool: (i) validity, (ii) reliability, and (iii) practicability.

Validity refers to the extent to which a test measures what it intends to measure. The measuring instrument should satisfy for its content validity, face validity, criterion validity and construct validity. The reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept and thus helps to access the goodness of a measure. A measure is said to process stability if one can secure consistent results with repeated measurements of the same subject (respondent) with the same instrument. Two tests of stability are test-retest reliability and parallel-form reliability. The internal consistency of measures is indicative of the homogeneity of items that measure the construct. The inter-item consistency reliability and split-half reliability are often used to test the internal consistency of measures. Finally, from the operational point of view, the measuring instrument ought to be economical, convenient and interpretable.

Discussion Questions

- 1 What are the essential differences among nominal, ordinal, interval and ratio scales? Discuss in terms of their properties with suitable examples.
- 2 Suppose you have the data from a corporation on the annual salary of each of its 300 employees.
 - (a) Illustrate how the data can be presented as ratio, interval, ordinal and nominal data.
 - (b) Describe the successive loss of information as the presentation changes from ratio to nominal scale.
- 3 Construct the Stapel's scale to access the counter service of TESCO supermarket.
- 4 You have been asked by the head of marketing to design an instrument by which any private university in Malaysia can evaluate the quality and value of its various curricula and courses. How will you ensure that your instrument has:
 - (a) stability?
 - (b) internal consistency?
 - (c) content validity?
 - (d) construct validity?
- 5 A valid instrument is always reliable, but a reliable instrument may not always be valid. Comment on this statement.



CHAPTER

8

Questionnaire Design and Fieldwork Plan

Learning Outcomes

After reading this chapter, student should be able to:

- Understand the basic rules of questionnaire design
- Implement do's and don'ts while designing the questionnaire
- Know how to validate the survey questionnaire
- Know how to pre-test the questionnaire
- Understand how to plan for fieldwork for survey

8.0 Introduction

The design of a questionnaire depends on the purpose of data collection. It depends on whether the researcher's objective is to collect qualitative information or quantitative information. The qualitative information is needed if the study is more exploratory in nature for the purpose of better understanding of a given research problem or the generation of hypotheses on a subject. The quantitative information is needed if the study is more of conclusive in nature which requires testing of hypotheses that have been previously generated.

Exploratory questionnaires: It may not be necessary to have formal questionnaire if the data to be collected is qualitative in nature and thus, does not require statistical validation. For example, a researcher might be interested to find out how decisions are made in the family

■ The formal questionnaire may not be necessary if the data requirement is qualitative in nature and it requires no statistical validation.

while purchasing foodstuffs for the breakfast. A formal questionnaire in this case may restrict the discussion and prevent complete exploration of views while interviewing the target respondent (i.e., decision maker in the household). Instead of designing the questionnaire, the researcher might prepare a brief guide, listing of important open-ended questions with appropriate probes/prompts listed under each

Formal standardized questionnaires: If the researcher is interested to statistically analyse the data to test the hypotheses for a conclusive study, a formal standardized questionnaire must be designed. Such questionnaires are generally characterized by:

- Prescribed wording and order of questions, to ensure that each respondent receives the same stimuli.
- Prescribed definitions or explanations for each question, to ensure interviewers handle questions consistently and can answer respondents' requests for clarification if they occur.
- Prescribed response format, to enable rapid completion of the questionnaire during the interviewing process.

8.1 Questionnaire Design for Business Research

If the same study with exactly same hypotheses is being set for five different researchers, most likely each one of them may come up with their own unique questionnaire. All the five questionnaires set by five researchers independently may differ widely in their choice of questions, wording, coding, sequencing, use of open-ended questions and so on. Though there is no hard and fast rule on how to design a questionnaire, but there are a number of points that should be borne in mind.

- A questionnaire should be designed in such a way that it meets the objectives of the study set by the researcher. Though it seems obvious, but a researcher may omit important aspects due to inadequate preparatory work or poor understanding of research problem. To some extent this is inevitable. Every survey is bound to leave some questions unanswered and provide a need for further research but the objective of good questionnaire design is to "minimize" these problems.
- The questionnaire design should be such that it obtains the complete and accurate information as far as possible. The researcher must ensure that questions are asked in such a way that respondent fully understand the meaning of the questions and are not likely to refuse to answer or lie to interviewer. A well-designed questionnaire is organized and worded to encourage respondents to provide accurate, unbiased and complete information.
- A well-designed questionnaire should ensure that it is easy for the respondents to respond to questions as well as easy for the interviewer to record the answer of the respondents.

- The questionnaire should be designed in such a way that it makes the interview concise and to the point, so as to make respondents remain interested till the end of the survey.

■ The questionnaire design should ensure that it is concise, easy to handle and obtains the complete and accurate information to meet the objectives set by the researcher

8.1.1 Defining the Target Population

The researcher must define the target population about which he/she wishes to generalize from the sample data to be collected. Secondly, researchers have to draw up a sampling frame, method of sampling and the sample size. Thirdly, in designing the questionnaire we must take into account demographic information such as the age, education, etc. of the target respondents to ensure that the characteristics of the sample are as similar as the characteristics of the target population.

8.1.2 Language

While designing the questionnaire, the first and foremost question you need to ask yourself as a researcher is "What language is the respondent going to understand and respond in?" The questionnaire should be designed in such a way that it can be used in any language. This does not mean it has to be printed in each language in which it has to be administered. For instance, a questionnaire printed in English could be administered to the respondent in the local language he/she speaks by a trained interviewer who could translate each question on the spot.

8.1.3 Deciding on What to Ask

The next obvious question is "What you are going to ask? In order to meet the objectives of the survey, it is important to decide what information one needs to collect from the respondents. A number of questions might come in one's mind related to the research problem that under investigation. But here the question is what questions should be asked and what questions should be avoided. One approach could be to ask oneself, "Is this question really needed?" The answer is YES, if this question produces the data which helps to test one or more of the hypotheses established during the research design. Generally, there is a high temptation in researchers to include the questions without critically evaluating their contribution towards the achievement of the research objectives. Researcher should control this temptation and avoid asking a particular question if it does not help in meeting any of the research objectives.

There are three potential types of information, researchers should be interested in:

- Information they are primarily interested in, that is, dependent variables.
- Information which might explain the dependent variables, that is, independent variables.
- Other factors related to both dependent and independent factors which may distort the results have to be adjusted for, that is, confounding variables.

Let us take as an example a national survey, to find out students' factors of predicting the level of certain knowledge, skills and attitudes at the end of

their undergraduate course. The dependent factors include the students' level of relevant knowledge, skills and attitudes. The independent factors might include students' learning styles, CGPA grades, socioeconomic status, ethnicity, etc. Confounding variables might include the types and quality of teaching in each undergraduate college/university.

8.1.4 Deciding How to Ask

Questions can be asked in the close-ended form or open-ended form. An open-ended question is one in which there are no standard answers to choose from. For example:

- 1 How old are you? _____ years.
- 2 What do you like best about your job?

On the other hand, a closed-ended question is one in which you provide the response categories, and the respondent just chooses one:

- 1 How old are you?
 - (a) 12–15 years
 - (b) 16–25 years
 - (c) 26–35 years
 - (d) 36–45 years
 - (e) Above 45 years
- 2 What do you like best about your job?
 - (a) The people
 - (b) The diversity of skills you need to do it
 - (c) The pay and/or benefits
 - (d) Other: _____ (specify)

There are lot of reasons for choosing one form over the other. Table 8.1 lists the advantages and disadvantages of close-ended and open-ended questions.

Table 8.1

Advantages and disadvantages of close-ended and open-ended questions

	Advantages	Disadvantages
Closed-ended	• Easy and quick to answer • Easy to compare responses across the respondents • Answers easier to analyse on computer • Response choices make questions clearer • Easy to replicate study	• Can put ideas in respondent's head • Respondents can feel constrained/frustrated • Many choices can be confusing • Cannot tell if respondent misinterpreted the question • Fine distinctions may be lost
Open-ended	• Permit unlimited number of answers • Respondents can qualify and clarify responses • Can find the unanticipated answer • Reveal respondents thinking process	• Answers can be irrelevant • Inarticulate or forgetful respondents are at disadvantage • Coding responses are subjective and tedious • Requires more response time and effort

8.1.5 Wording of Individual Questions

The way questions are phrased is important and there are some general rules for constructing good questions in a questionnaire.

Use short and simple sentences

Short, simple sentences are generally less confusing and ambiguous than long, complex ones. As a rule of thumb, most sentences should contain one or two clauses. Sentences with more than three clauses should be rephrased. There is every chance that some of your target respondents may have a low level of education and thus they may find it difficult to understand the long and complex questions completely. The researcher may not have an idea how well and badly educated the target respondents are until he/she gets into the fieldwork. So it is always a good practice to keep the question as simple and short as possible.

Ask for only one piece of information at a time

A question should look for only one piece of information at a time. If two informations are included in one question, it is referred as to a *double-barrelled question* and double-barrelled questions should always be avoided.

Spot the difference

- (a) Please rate the lecture in terms of its content and presentation.
(i) very bad (ii) bad (iii) neither good nor bad (iv) good (v) very good
- (b) Please rate the lecture in terms of its content.
(i) very bad (ii) bad (iii) neither good nor bad (iv) good (v) very good
- (c) Please rate the lecture in terms of its presentation.
(i) very bad (ii) bad (iii) neither good nor bad (iv) good (v) very good

In the above examples, (a) is a double-barrelled question as it looks for two pieces of information—content as well as presentation of lecture. If both the information is necessary, the right approach is to ask two different questions as (b) and (c).

Avoid negatives if possible

Try to ask positive question as far as possible. For example, instead of asking students whether they agree with the statement, "Small-group teaching should not be abolished," the statement should be rephrased as, "Small-group teaching should continue". Double negatives should always be avoided.

Spot the difference

- (a) Are you not happy with the teaching of BRM course?
(i) yes (ii) no
- (b) Are you happy with the teaching of BRM course?
(i) yes (ii) no

In the above examples, (a) is a negative question. If student replies "Yes", it actually means "No" (he/she is not happy). If student replies "No", it actually means "Yes" (he/she is happy). As such there is no problem with the question, but there is every chance that a casual respondent may reply against his own perception. So, the right way to ask the question is (b) where reply to "Yes" means yes and reply to "No" means no.

■ The questions should be short, simple and precise. One should avoid asking negative question or more than one piece of information at a time.

Ask precise questions

Questions may be ambiguous because a word or term may have a different meaning. For example, if we ask students to rate their interest in "medicine", this term might mean "general medicine" (as opposed to general surgery) to some, but inclusive of all clinical specialties (as opposed to professions outside medicine) to others. Another source of ambiguity is a failure to specify a frame of reference. For example, in the question, "How often do you borrow books from your library?" the time reference is missing. It might be rephrased as, "How many books have you borrowed from the library within the past six months altogether?" Similarly, the question "what is your income" is vague as different respondents may interpret this question in different dimensions such as "hourly pay", "weekly pay", "yearly pay", "income before tax", "income after tax", "personal income", "family income" and so on. The researcher needs to clearly specify the "term" within which the respondent has to answer so as to avoid any personal interpretation of question by the respondent.

Questions must generate the required information

The number of questions to be asked must be good enough to generate the required information for the study. In a taste panel exercise, just asking a question "which product do you prefer?" will certainly neither reveal anything about attribute(s) the product was judged upon nor reveal the degree of preference. In such cases a series of questions would be more appropriate.

Words must have the same meaning to all respondents

Spot the difference

- (a) "How many members are there in your family?"
- (b) "How many siblings do you have?"

■ The questions must generate required information and convey the same meaning to all the respondents.

In question (a) there is room for ambiguity since it is open to interpretation as to whether one is speaking of the immediate or extended family. The question (b) will mean the same thing (how many brothers and sisters including him/her) to all the respondents.

Avoid leading questions

While framing the question you have to see if any of the words or phrases loaded or leading in any way.

Spot the difference

- (a) What did you dislike about the product you have just tried?"
- (b) (i) "Did you dislike any aspect of the product you have just tried?"
Yes () No () (If No, answer (ii))
(ii) "What did you dislike about the product?"

In question (a) it is assumed that all the respondents dislike the product. They are not given an opportunity to indicate that there is nothing he/she dislikes about the product. To avoid the biasness in a question, a preliminary question can be asked as in (b).

Spot the difference

- (a) "You would not say that you were in favour of school on Saturday morning, would you?"
- (b) "Would you say that you are not in favour of school on Saturday morning?"
- (c) "Do you favour or oppose school on Saturday morning?"

The question given in (a) or (b) is an example of leading question. The right way of asking the question is as given in (c).

Avoid hypothetical questions

As far as possible you should avoid asking a hypothetical question. Hypothetical questions such as "would you use this resource in your class room teaching if it were available?" are not considered good for behavioural prediction. It is generally difficult for people to predict their own behaviour because of changing circumstances as a result of interference of various situational factors surrounding them. Data is more valid if the questions are about the past behaviour and present circumstances, attitudes and opinions.

Do not overtax the respondent's memory

There is a high probability of getting poor quality of information if the respondents are asked to recall past behaviour over a long retrospective period. This is true especially when recurrent events or behaviours are concerned. For example, it is difficult for a student to remember "how many hours of internet browsing he/she could do on an average in the last month". The reliability of the response to such questions may be questionable because the time could be just too long to remember what happened in details. If at all, such question need to be asked to meet research objective, a one-week recall period might be more appropriate.

■ The leading questions, hypothetical questions or the questions that demand respondents to recall past behaviour over a long retrospective period should be avoided.

Ensure those you ask have the necessary knowledge

Before framing the question, make sure that respondents have the necessary knowledge to answer it. For example, in a survey of university lecturers on recent changes in higher education, the question, "Do you agree with the recommendations in the Dearing report on higher education?" is unsatisfactory for several reasons. Not only does it ask for several pieces of information at the same time as there are several recommendations in the report, the question also assumes that all lecturers know about the relevant recommendations.

Level of details

It is important to ask for the exact level of details required. On the one hand, you might not be able to fulfil the purpose of the survey if you omit to ask essential details. On the other hand, it is important to avoid unnecessary details. People are less inclined to complete long questionnaires. This is particularly important for confidential sensitive information, such as personal financial matters or marital relationship issues.

Sensitive issues

It is often difficult to obtain truthful answers to sensitive questions. Clearly, the question, "Have you ever copied other students' answers in a degree exam?" is likely to produce either no response or negative responses. Less direct approaches can be used when a question is sensitive to the respondents. Some of the approaches that can be used to ask the question are given below.

Casual approach: "By the way, do you happen to have copied other students' answers in a degree exam?" may be used as a last part of another decoy question.

Numbered cards approach: "Please tick one or more of the following items which correspond to how you have answered degree examination questions in the past." In the list of items, include "copy from other students" as one of many items.

Everybody approach: "As we all know, most medical students have copied other students' answers in degree exams. Do you happen to be one of them?"

Other people approach: This approach was used in the recent medical student survey. In this survey, students were given the scenario, "John copies answer in a degree exam from Jean". They were then asked, "Do you feel John is wrong, what penalty should be imposed for John, and have you done or would you consider doing the above?"

The question asked above in four approaches is indirect and relatively less sensitive. So the chances of revealing the true response by respondents are much higher as compared to when the question is asked directly.

■ If a question is found to be sensitive, the question can be asked in a less direct way to have higher chances of getting the true response from the respondents.

Minimize bias

People tend to answer questions in a way they perceive to be socially desired or expected by the questioner, and they often look for clues in the questions. Many apparently neutral questions can potentially lead to bias. For example, in the question, "Within the past month, how many lectures have you missed due to your evening job?" students may perceive the desired responses to be "never" to the first question. This question could be rephrased as, "Within the past month, how many times did your evening job commitment clash with lectures? How many times did you give priority to your evening job?" Take another example. The question, "Please rate how useful the following textbooks are. Please also state whether they are included in your lecturer's recommended reading list?" There is a risk that students may perceive that they should rate books recommended by lecturers more favourably than those not recommended by their lecturers. This risk may be minimized by putting the second question later on in the questionnaire.

8.1.6 Sequencing of Questions*Opening questions*

The questionnaire should begin with easy and non-threatening question. Initial few questions are crucial because it is respondent's first exposure to interview

and sets the idea on the nature of task to be performed. If the initial questions are difficult to understand, or beyond their knowledge and experience, or embarrassing in some way, there is every chance that respondents may be discouraged to actively participate in a survey or completely deny to participate.

Question flow

Questions should be asked in the sequence of a psychological order so that one leads easily and naturally to the next. In order to make respondents to feel connected, questions on one subject or one particular aspects of a subject should be grouped together.

■ The questionnaire should begin with easy and non-threatening questions and follow the sequence of a psychological order so that one leads easily and naturally to the next

Question variety

Respondents may feel monotonous when similar questions with the same pattern are asked for half an hour and so. An open-ended question here and there (even if it is not analyzed) may provide much-needed relief from a long series of questions in which respondents have been forced to limit their replies to pre-coded categories. Questions involving showing cards/pictures to respondents can help vary the pace and increase interest.

In summary, following points should be considered while sequencing questions in the questionnaire.

- Put the most important items in the first half of the questionnaire.
- Do not start with awkward or embarrassing questions.
- Start with easy and non-threatening questions.
- Go from the general to the particular.
- Go from factual to abstract questions.
- Go from closed to open questions.
- Leave demographic and personal questions until last.

8.1.7 Length of Questionnaire

There are no universal agreements about the optimal length of questionnaires. It probably depends on a number of factors such as the number of objectives in the study, type of respondents (whether target respondents are consumers, managers, students or kids) or type of survey (whether it is a personal interview, postal survey or internet survey). As a rule of thumb, ask only necessary questions and avoid unnecessary one. You need to ask yourself if a question is going to help in meeting any of the objectives set out for the research. If the answer is NO, you should avoid asking that particular question.

8.1.8 Ease of Recording

Questionnaire design should ensure it is easy to carry, visible in different kinds of light, and the distance between different answer categories should be sufficient so that there is no confusion or mistake while placing a tick over the actual response or a given question.

8.1.9 Coding

If the questionnaire is coded before doing the fieldwork (as most of the questionnaire are likely to be these days), it must be ensured that the fieldworkers know where to mark the answer – on the code or actual answer choice.

Spot the difference

In a typical day how often do you smoke?

Format 1:

Never Occasionally ☒ Sometimes Often Regularly

Format 2:

Never__Occasionally ☒ Sometimes__Often__Regularly__

Format 3:

☐Never__Occasionally ☒ Sometimes__Often__Regularly

Format 1 is not coded and thus it may lead error either because of placing tick (☒) at the wrong place by the respondent or reading the response wrongly by fieldworker. It is not very clear from the response whether the respondent smoke occasionally or sometime. So, question must be properly coded to avoid any confusion, such as Format 2 (or Format 3) where respondents are expected to put (☒) on the right side (or on the left side) of the option.

8.1.10 Analysis Required

It is important to plan for analysis well before designing the questionnaire. Regrettably, this is not always given due attention by the researchers. It is sometimes assumed that it can be done later, or that all possible analysis can be thought of after collecting data, so why to bother to plan the analysis in advance. But for many reasons, it is vital to do so. How you are going to ask the question (whether open-ended or close-ended) and what is the measurement scale (for example, rating or raking) depend very much on how you are going to analyse the data. For example, to run statistical technique multidimensional scaling, the questionnaire is to be constructed in a particular way. Similarly if you plan to analyse the factor analysis to empirically find out important dimensions of a theoretical construct, the questions need to be measured on a rating scale.

8.2 General Appearance of a Questionnaire

The physical appearance of a questionnaire can have a significant effect upon both the quantity and quality of survey data obtained. The appearance of questionnaire has an impact on respondents' motivation in giving response. Always provide enough space for respondents to answer the questions and sufficient space between two questions so that the questionnaire does not look clumsy. Use clear headings and numberings wherever appropriate. Generally, it is found that researchers often use smaller fonts and very little spacing between the questions in order to reduce the number of pages. One

needs to be careful as use of very small font size may result in an unreadable questionnaire. In no case, less than 10 point text size should be used.

8.2.1 Introduction to Respondents

It is important to have the covering letter if the questionnaire is to be mailed or distributed to respondent. The purpose of such a letter is to introduce the respondents with the objectives of the survey in order to encourage them to actively participate in the survey. In an interview, one of the tasks of the interviewer is to persuade the respondents to cooperate. However, in the self-administered questionnaire, the covering letter is the only instrument for overcoming resistance. The covering letter should include the following:

■ It is important to have the covering letter if the questionnaire is to be mailed or distributed to respondents.

- Name of the organization (such as Ministry of Education, University, Marketing Research Company, etc.).
- Purpose of the study.
- Declaration that ensures the respondents that the information provided will be managed in a strictly confidential manner and that all respondents will remain unidentified.
- Explain how important is their involvement in actively and genuinely participation in the survey.
- Name and contact numbers of the Principal Researcher.

The following additional information should also be included in both the introduction to the questionnaire and the covering letter:

- Brief detail on how the respondent was selected (for example, "your name was randomly selected").
- Expression of appreciation for the respondent's help.
- Estimate of questionnaire completion time.

EVALUATING SERVICE QUALITY OF BANKS IN MALAYSIA

Dear Bank Customer

This survey is part of management project entitled "Evaluating Service Quality of Banks in Malaysia" for the partial fulfilment of my degree in the Faculty of, University....., Malaysia. As a bank consumer, your views on banks' service quality are invaluable in helping me in completing this project. The survey will collect and assess the *expectations* and *perceptions* of banks' service quality, including *profile* of the consumers, their *banking behaviour* and *preferences*. Your active participation and genuine response will be highly appreciated. Hereby, I assure you that the information collected from you will not be disclosed to third party and used only for my research work.

Thank you,

(XXXX)

UNDERSTANDING CONSUMERS' ORIENTATION AND POST-PURCHASE BEHAVIOUR IN FINE DINING RESTAURANT

Dear Restaurant Guest

I am a Ph.D. student at the Faculty of, University, Malaysia. For my doctoral dissertation, I am conducting a study to determine fine dining restaurant consumers' purchasing orientation (active or passive consumers) and their post-purchase behaviour (perception and satisfaction with the restaurant as well as the outcome of their dining experience). Your kind assistance is absolutely vital to the success of this study. I shall be grateful if you could spare about 10 to 15 minutes of your valuable time in filling up the questionnaire that will be distributed to you by the restaurant's staff and collected by them before you leave the restaurant. You are not required to provide your name for this study and I assure you that your responses will only be used for academic purposes.

Thank you

(XXXX)

8.2.2 Instructions

Interviewer instructions should be placed alongside the questions to which they pertain. Instructions on where the interviewers should probe for more information or how replies should be recorded are placed after the question.

INSTRUCTIONS TO A RESPONDENT FOR A SET OF QUESTIONS

Instruction: The following statements relate to your feelings about the particular bank you frequently do your banking transactions with. Please show the extent to which you believe the bank has the feature described in the statement..

INSTRUCTIONS TO A RESPONDENT FOR A SET OF QUESTIONS

Please indicate how strongly you agree or disagree on a set of statements on online banking on the following Likert scale. Please circle the answer.

1= Strongly disagree, 2= Disagree, 3= Somewhat disagree, 4= Neither agree nor disagree, 5= Somewhat agree, 6= Agree, 7= Strongly agree

INSTRUCTIONS TO A RESPONDENT FOR A SINGLE QUESTION

Instruction: About how many different teaching positions have you held during your life? (Count only those teaching positions that you have held for at least one full academic year.)

8.2.3 Concluding Questionnaire with an Open-ended Question and Thanks

There is every likely the case that a respondent become increasingly indifferent to the questionnaire as it nears the end and thus may become impatient or fatigue. This in turn may result in careless answers to the later questions. The questions that are of special importance should be placed in the earlier part of the questionnaire. Sensitive questions should be asked towards the end to avoid the respondents cutting off the interview before the information is collected. It is always good to end the questionnaire with an open-ended question to get the free opinion on the topic. At the end of the questionnaire do not forget to "thank" the respondent once again for their valuable time spent on completing the survey.

■ The questions that are of special importance should be placed in the earlier part of the questionnaire and the sensitive questions towards the end of the questionnaire.

8.3 Pre-testing Questionnaire

Pre-testing the questionnaire is an essential step before its completion. The purpose of the pre-test is to check question wording, and to obtain information on open-ended questions with a view to design a multiple choice format in the final questionnaire. The purpose of pre-testing the questionnaire is to determine:

- Wording of the questions are correct to convey the same meaning to all the respondents.
- Whether the questions have been placed in the right sequence.
- Whether the questions are clearly understood by all classes of respondents.
- Whether additional questions are needed or whether some questions should be eliminated.
- Whether the instructions to interviewers are clear and adequate.

Usually, the respondents selected for pre-test need not be large. However, the respondents selected for the pilot survey should be broadly representative of the sample to be chosen for the main study. Note that questions "borrowed" from existing questionnaires need to be pre-tested to ensure that they will work as required with the "new" respondents. The first version of the pre-test questionnaire often contains considerably more questions than the final questionnaire. This can be upsetting for the respondents – especially if many questions are asked in an unstructured and open form so that the amount of time required to complete the questionnaire is considerable. If absolutely necessary, the questionnaire could be divided in two or three parts (of equal length and answering time) for the first tryout, so that each respondent answers only a fraction of the questions. For each form at least 50 respondents should be asked to participate. The information collected in this first pre-test should provide sufficient information to produce a second version of the questionnaire for final pre-testing. This second version of the questionnaire will then be administered in one single form in order to further verify the functioning of the items and answer categories, as well as that of the questionnaire overall structure, layout and accompanying instructions.

■ The fieldwork plan should be well planned to ensure timely completion of data collection.

8.4 Design of Fieldwork

Careful planning is required for prompt receipt of survey questionnaires to the different areas covered in the data collection undertaking. A clear plan for data collection should be developed by the researcher so that:

- there is a clear overview of what tasks need to be carried out, who should perform them and what should be the duration of these tasks;
- human and material resources for data collection are organized in the most efficient way; and
- delays in data collection because of lack of planning are minimized.

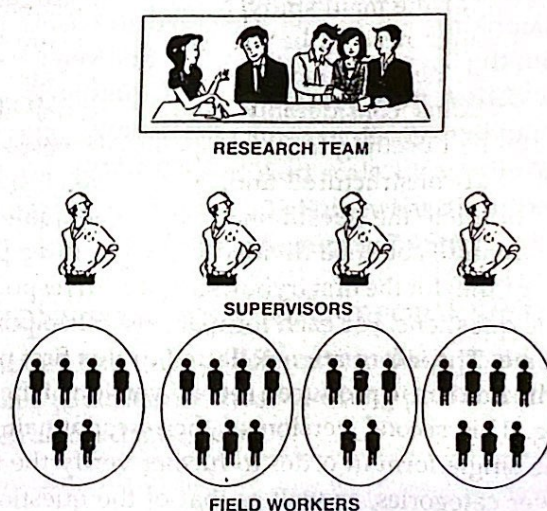
8.4.1 Fieldwork Plan

Fieldwork plan is clearly linked to the sampling plan. Once the sampling area (cities, town, etc.) and the sample size are determined for each, the next step is to plan on the following:

- *Who will do the fieldwork?*
- *When should the fieldwork start?*
- *How long should the fieldwork be carried on?*

First, you need to plan who will do the fieldwork. Fieldwork assumes that we are collecting data from the target respondents by going to the field. The field is any place (homes, offices, shopping mall, restaurants, etc.) where the fieldworkers find the respondents to conveniently provide information. In a typical research project where huge data is to be collected from wider geographical area within the targeted time period, the research team appoints the supervisors at different parts of the geographical area. Each supervisor appoints the fieldworkers under him/her whose responsibility is to monitor the fieldworkers from time to time and report on the progress of data collection to the research team (Figure 8.1). It is the fieldworkers who actually go to the field and collect information from the targeted respondents within his defined area.

Figure 8.1
Structure of the research team
in a fieldwork plan.



The second question is when the data collection should be carried out. Before we go for fieldwork, we make sure that the questionnaire is final. Once fieldworkers are on the field to do the survey, any changes or moderation in the questionnaire is not desirable. In many studies carried out nationally, it is not possible to simultaneously cover the entire centre on the same day. There could be logistic problems for supervisors, or there may be difficulty in recruiting adequate fieldworkers and so on. But it is desirable to have a well-planned schedule so that all fieldwork is completed in an orderly fashion.

For the third issue, i.e. for the time requirement for data collection, one needs to consider the following points.

Step 1: Calculate the:

- time required to reach the study area(s),
- time required to locate the study units (persons, groups or records), and
- number of visits required per study unit. In the longitudinal studies, the survey needs to be carried out from time to time, whereas in the cross-sectional study, the survey is done once forever.

Step 2: Calculate the number of interviews that can be carried out per fieldworker per day.

Step 3: Calculate the number of days needed to carry out the interviews. For example:

- you need to do 200 interviews,
- your research team of 5 people can do $5 \times 4 = 20$ interviews per day,
- you will need $200:20 = 10$ days for the interviews.

8.4.2 Briefing and Debriefing

It is the responsibility of the research executive in charge to personally brief the field supervisors who are responsible for supervising the team of fieldworkers during the period of data collection. This briefing session is conducted after the recruitment of fieldworkers, and ends with a practical round of mock interviews and questions from fieldworkers on the possibility of any difficulties they may face such as in identifying and locating target respondents, asking certain questions and so on. The objective of mock interview and the briefing session is to explain and clarify the fieldworkers on how to complete the data collection task effectively. During the time of data collection, the fieldworkers may encounter a number of problems such as difficulty in locating target sample units or non-cooperation in answering some questions or difficulty in comprehension. If any such problem is reported to the field supervisor or the researcher, immediate solution is to be found out. To minimize any problems the fieldworkers may encounter, a debriefing session is usually held at the end of the first day of fieldwork in each location.

Possible solutions are thought of by the research executive and supervisor, and implemented for the remaining part of the study.

Chapter Summary

For a successful survey, a well-designed questionnaire is essential. Designing questionnaire is an art and there is no theory of the questionnaire to guide the researchers. So the researcher must develop through experience his/her own intuition with respect to what constitutes good design. A good questionnaire is one which helps directly to achieve the research objectives, provides complete and accurate information, is easy for both interviewers and respondents to complete, and is so designed as to make sound analysis and interpretation possible.

There are some basic rules that should be taken into consideration while designing the questionnaire. The researchers should carefully take the decision on a number of elements such as the language, deciding what and how to ask, wording of an individual question, sequencing of questions, length of questionnaire, coding and recording, and analysis required. The physical appearance of a questionnaire is very important as it can have a significant

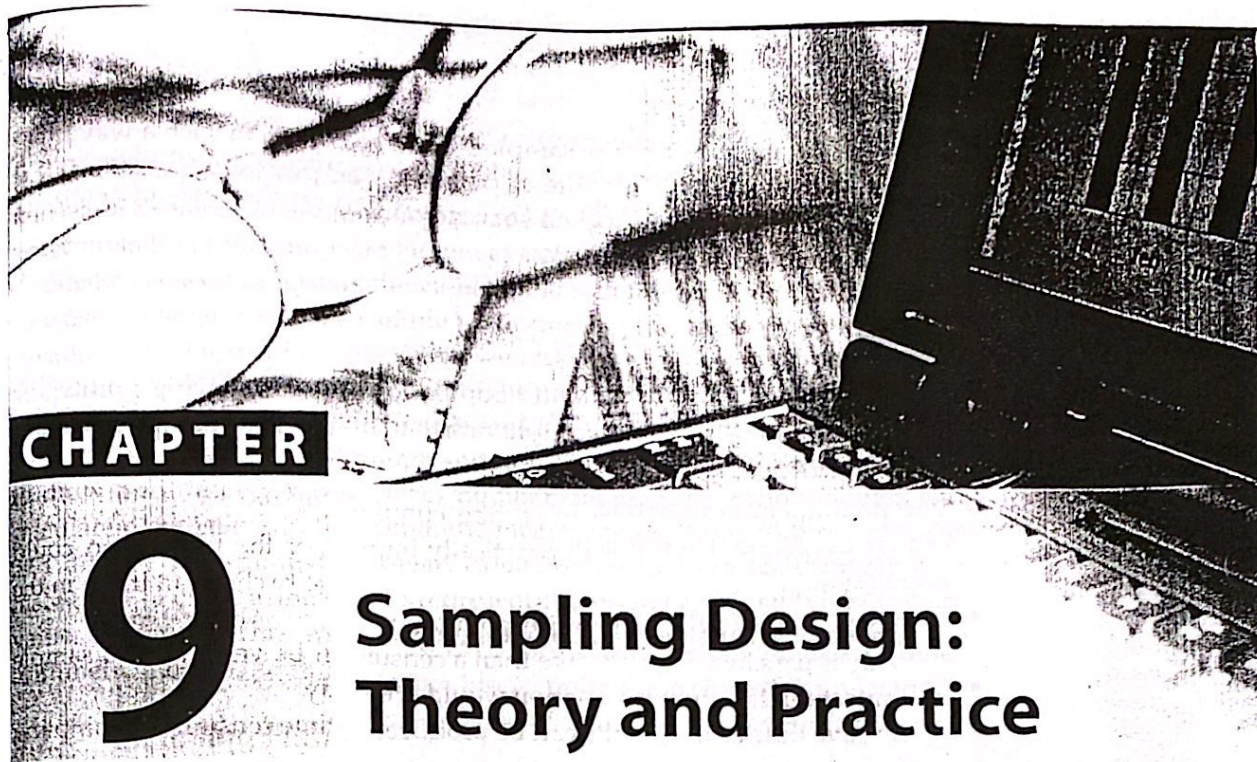
effect upon both the quantity and quality of survey data obtained.

Pre-testing the questionnaire is an essential step before its completion. The purpose of the pre-test is to check question wording, and to obtain information on open-ended questions with a view to designing a multiple choice format in the final questionnaire. Further, we should ensure the validity of the instruments—the degree to which a question measures what it intends to measure.

Once the survey instrument is ready, the researcher team need to plan for fieldwork so as to complete the survey effectively in time. In a typical research project where huge data is to be collected from wider geographical area within the targeted time period the research team need to appoint the supervisors at different locations whose responsibility is basically to monitor the fieldworkers and report progress on data collection to the research team.

Discussion Questions

- 1 Write down the advantages and disadvantages of open- and close-ended questions.
- 2 What are the basic rules that one should follow while wording an individual question in the questionnaire?
- 3 Discuss the various approaches (with an example) to handle sensitive questions in a questionnaire.
- 4 How important is the physical appearance of the questionnaire? Discuss the points that should be considered for better appearance of the questionnaire.
- 5 In a typical sponsored project, what is the structure of a team in a fieldwork plan? Discuss the role of researcher and field supervisor.



CHAPTER

9

Sampling Design: Theory and Practice

Learning Outcomes

After reading this chapter, student should be able to:

- Understand why you need to sample the population
- Know the basic terminology
- Understand the differences between probability and non-probability sampling
- Apply the appropriate sampling technique
- Determine the sample size
- Understand the factors that could affect the sample size in any study
- Understand the different types of errors in research

9.0

Introduction

The concept of sampling is used in our day to day life. One taste of a drink tells us whether it is sweet or sour. Similarly, by seeing and touching handful rice, we judge about the whole bag of rice. If some members of staff favour a promotional activity, we infer that others will also. These examples vary in their representativeness, but each is a sample.

Sampling may be defined as the selection of some part of the population on the basis of which a judgement or inference about the entire population is made. In most of the research work and surveys, the usual approach is to make generalizations or to draw inferences based on samples about the parameters of the population from which

the samples are taken. So the sample should be drawn in such a way that it is true representative of the entire population. The process of sample selection is called sample design and the survey conducted on the basis of sample is described as sample survey.

9.1 Why Sampling?

The question is "Why sample the population?" The following points justify why we need to choose the sample rather than going for complete census of the target population.

- The population is dynamic, i.e. the component of the population could change over time. Thus, it is practically impossible to check all items in the population.
- The cost of studying the entire population could be very high. A sample study is usually less expensive than a census study.
- Contacting the whole population would often be time consuming. Sampling can save time as the results can be produced at a relatively faster speed.

Sampling remains the only way when population contains infinitely many members or when the experiment involves the destruction of the items under study.

Sampling usually enables to estimate the sampling errors and thus, assists in obtaining information concerning some characteristics of the population.

9.2 Defining Basic Terminology

Before we go further in depth into the issue of sampling, let us first define some of the terms which will be used hereafter.

9.2.1 Population, Element and Population Size

Population refers to the entire group of people, events or things of interest that researcher wishes to investigate. For instance, if the marketing manager of Samsung is interested to know the kind of advertising strategies adopted by electronic companies in Malaysia, then all the electronic companies situated in Malaysia form the population. Similarly, if a researcher is interested to measure the service quality of Malaysian Islamic banks from the customers' perspective, then the population consists of all the Islamic bank's customers in Malaysia.

Each member of the population is known as *element*. The total number of elements in the population is known as *population size* and it is denoted by "N".

9.2.2 Sample, Subject and Sample Size

The *sample* is the subset of the population. Each member of the sample is known as *subject*. The total number of subjects in the sample is known as *sample size* and it is denoted by "n".

9.2.3 Parameter and Statistics

The characteristics of the population are known as *parameters* whereas the characteristics of the sample are known as *statistics*.

Statistic is used to estimate the value of the parameter. Note that the value of statistic changes from one sample to the next which leads to a study of the *sampling distribution of statistic*. When we draw a sample from a population, it is just one of many samples that might have been drawn and, therefore, observations made on any one sample are likely to be different from the "true value" in the population (although some will be the same). Imagine we were to draw an infinite (or very large) number of samples of individuals and calculate a statistic, say the arithmetic mean, on each one of these samples and that we then plotted the mean value obtained from each sample on a histogram (a chart using bars to represent the number of times a particular value occurred). This would represent the sampling distribution of the arithmetic mean.

9.2.4 Sampling Frame

It is a complete listing of the population of interest from which the sample is drawn. All members of the sampling frame have a probability of being selected. Without some form of *sample frame*, a random sample of a population, other than an extremely small population, is impossible.

When a list of the population of interest is not available, an alternate method for capturing the population must be found. Most surveys carried out by governmental statistical agencies rely on a sample frame that is composed of maps that partition the entire country into *enumeration areas*. In that case, a multistage sample design is required. Enumeration areas are first randomly sampled, and then individual housing units are sampled from within the enumeration areas. Finally, individuals are sampled from within the housing units.

Even though the set of maps of enumeration areas is not a list of individuals in the population, it is still considered a sample frame. In that case, however, it is a sample frame of individuals that reside in housing units, not of the total population. Any individual who does not live in a housing unit, for example, a homeless person, is not covered by the sample frame.

9.3 Sampling Techniques

Sampling is the process of selecting a sufficient number of elements from the population, so that a study of the sample and understanding of its properties or characteristics would make it possible for us to generalize such properties or characteristics to the population elements. In the process of sampling, we are selecting some elements of the population as the subjects of the sample. Which elements of the population are being selected as the subjects depend on the sampling technique that is chosen by the researcher? The sampling technique can be broadly divided into probability sampling and non-probability sampling. Each of these two broad categories consists of a number of sampling

techniques. The selection of a particular sampling technique depends on a number of factors such as the extent of generalizability desired, the availability of time and other resources, and purpose of the study.

9.3.1 Probability Sampling Techniques

In probability sampling technique, a sample is being selected using random selection so that each element of the population has a known chance of being selected. It is generally assumed that a *representative sample* is more likely to be the outcome when this method of selection from the target population is employed. Thus, findings based on probability sampling can be generalized to the target population with a specified level of confidence. The most commonly employed method of probability samplings are discussed below.

Simple random sampling

■ In simple random sampling, each and every member of the population has an equal and known chance of being the subject of the sample.

The simple random sampling is the most basic form of probability sampling. Here the sample is drawn from the target population in such a way that each and every member of the population has an equal and known chance of being the subject of the sample. The selection of each unit is independent of the selection of every other unit. Selection of one unit does not affect the chances of any other unit.

The procedure for drawing large sample involves the following steps:

1. Sequentially assign a unique identification number to each element of the population.
2. Use a random number generation (such as lottery method) to identify the appropriate elements to be the part of sample.
3. Ensure that no element is selected more than once.

Let us suppose, the HR manager of Management and Science University (MSU) wish to conduct a survey among the employees on their job satisfaction. Looking at the budget and time constraint, the HR manager wish to interview 20% of total 300 employees working in MSU. Thus, the sample size (n) chosen for the study is 20% of 300, i.e., $n = 60$. In this case each employee has the equal probability of $60/300$, i.e., 0.2 for being selected as the subject of the sample. The Manager can use the employees' ID number as the identification number to prepare 300 number of small tags, each tag containing one ID number. All tags can be placed in a bowl or a hat and mixed thoroughly. The blind-folded researcher then can pick numbered tags from the bowl one by one till he gets the required number of sample. Another way would be to let a computer do a random selection from the population. For populations with a small number of members, it is advisable to use the first method but if the population has many members, a computer-aided random selection is preferred. One can use SPSS (statistical package for social sciences) to generate random sampling. SPSS is a comprehensive system for analyzing data. It is available for both personal and mainframe computers.

Using SPSS to Select Random Sample

- 1 Go to Data → Select Cases
- 2 Select Random Sample of Cases and then Click on Sample in Select Cases Box
- 3 In Select Cases: Random Sample Box, choose first option Approximately [] of all cases or else, second option Exactly [] cases from the first [] cases
- 4 Click Continue
- 5 In Output window in Select Cases Box, choose second option Copy selected cases to a new dataset
- 6 Click OK

Stratified sampling

While sampling helps to estimate population parameters, there may be identifiable subgroups of elements within the population that may be expected to have different parameters on a variable of interest to the researcher. In stratified random sampling, the population is divided into different subgroups known as strata on the basis of some criteria. Then the method of simple random sampling is used to draw the sample from each stratum.

Let us suppose, the HR manager of company X is interested in assessing the level of job satisfaction among the employees. The researcher in this case may find it worth doing the study at different job levels rather than pulling all the employees together under one umbrella. This is so because the employees at the top management will have different criteria for their job satisfaction than those who are at the clerk level. Let us suppose, there are total 885 employees in the organization. In this case the whole population can be divided into different subgroups (strata) on the basis of job level as given in the table below.

■ The process by which the sample is drawn randomly from each mutually exclusive subgroup (stratum) is known as stratified random sampling

Job level	No. of elements	Proportionate sampling	Disproportionate sampling
Top management	$N_1 = 15$	$n_1 = 3$	$n_1 = 8$
Middle-level management	$N_2 = 40$	$n_2 = 8$	$n_2 = 17$
Lower-level management	$N_3 = 60$	$n_3 = 12$	$n_3 = 25$
Supervisors	$N_4 = 140$	$n_4 = 28$	$n_4 = 40$
Clerks	$N_5 = 600$	$n_5 = 120$	$n_5 = 75$
Secretaries	$N_6 = 30$	$n_6 = 6$	$n_6 = 12$
Total	$N = 885$	$n = 177$	$n = 177$

Column 1 shows the different subgroups (strata) followed by number of employees in each strata shown in column 2. Now the sample can be drawn from each stratum either proportionately or disproportionately. Suppose we have decided to take 20% of the population as our sample size. In proportionate stratified random sampling, the sample taken from each stratum will be equal to 20% of the population size of the strata as shown in column 3.

However, researcher might sometimes be concerned that the sample chosen based on proportionate sampling may result in insufficient number of sample for some of the strata. For example, in the above case a sample of just 3 members from the top-level management may not truly reflect how all members at those levels would respond. Therefore, a researcher might decide instead to

use a disproportionate stratified random sampling as shown in column 4 of the table. In disproportionate stratified random sampling, the percentage of population taken as the sample for different strata is not the same. The idea here is that 75 clerks might be considered adequate to represent the population of 600 clerks and at the same time, 8 out of 15 top-level managers would also be considered as representative of the top-level management. Disproportionate sampling decisions are made either when some stratum or strata are too small or too large, or when there is more variability suspected within a particular stratum. For example, the educational levels among supervisors, which may be considered as influencing perceptions, may range from elementary school to master's degrees. Thus, more samples may be required at the supervisor level. Disproportionate sampling is also sometimes done when it is easier, simpler, and less expensive to collect data from one or more strata than others.

Systematic sampling

■ In systematic sampling, every k th element in the population is sampled, beginning with a random start of an element in the range of 1 to k

Systematic sampling is very similar to random sampling, and is easier in practice. Once the sample size is decided, we divide the total population into n parts, where " n " is the sample size required. From the first part of sampling units, we pick up one at random. We then pick up every (N/n) th item from the remaining parts. Let us suppose for a study we require 15 households from a total population of 300 houses in a particular locality. The sampling fraction = $15/300$ which means 1 out of every 20 houses will be selected. On an average, we divide the list into n ($=15$) parts. Out of the first 20 houses, we chose any one house number at random. Let us say, the randomly picked up house is No. 7. Therefore we chose house numbers as 7, $7 + 20$, $7 + 20 + 20$, and so on in a systematic sampling plan.

Cluster sampling

A cluster is a group of sampling units or elements which can be identified or listed and a sample of which can be chosen. In simple random sampling, elements are randomly selected from the entire population of study. In cluster sampling, population is divided into groups of elements with some groups randomly selected for the study.

■ A cluster is a group of sampling units or elements which can be identified or listed and a sample of which can be chosen

However, it is different from stratified sampling. In stratified sampling, we take a subset of population sampling units within each stratum to form the sample. On the other hand, in cluster sampling, we take a subset of strata as the primary sampling units. When the strata themselves are the primary sampling units, the strata are called clusters. The selection of a sample of clusters to provide a sample of population units is called cluster sampling. If all of the population units in every selected cluster are in the sample, then this is known as one-stage cluster sampling. The two primary reasons cluster sampling is employed for sample surveys of human populations in large geographical regions are feasibility and economy. Cluster sampling is often the only feasible method of probability based sampling because the sampling frames for the target populations are lists of clusters. Further, it is economical compared to simple random sampling as the subjects of the sample are selected from the randomly selected strata rather than from the entire population. However, the

statistical efficiency of cluster sampling is usually lower than simple random sampling. The reason for this is that the listing units within the same cluster tend to be homogeneous with respect to many characteristics. Because of the homogeneity, the information collected from within a cluster tends to be redundant. For example, in a household survey, the people within a household have the same socioeconomic status, usually are of the same ethnic origin, hold many of the same beliefs and usually have the same dietary habits. People within a block of houses usually have the same tendencies depending on the neighbourhood. In statistical terms, there is a high positive correlation between attributes within the respondents in the same cluster.

Multistage cluster sampling

In single-stage cluster sampling, the population is divided into convenient clusters and we randomly choose the required number of clusters as sample subjects and investigate all the elements in each of the randomly chosen cluster. Cluster sampling can also be done in several stages and it is known as multistage cluster sampling. For example, it may not be possible to list all of the customers of a chain of GAIN Supermarket in Malaysia. However, it would be possible to randomly select a subset of malls based on say geographical area (stage 1 of cluster sampling) and then randomly select areas in each of these locations (stage 2 of cluster sampling)

■ When cluster sampling is done in several stages, it is known as multistage cluster sampling.

In a study to estimate the average bank deposits with banks at national level, the cluster sampling can be used in several stages:

- Cluster 1: Selection of geographical areas within Malaysia such as urban, semi-urban and rural.
- Cluster 2: Selection of particular areas in each of these locations.
- Cluster 3: Selection of banks within each selected area.

The multistage cluster sampling involves a probability sampling of the primary sampling units (Cluster 1); from each of these primary units, a probability sample of the secondary sampling units (Cluster 2) is then drawn; a third level of probability sampling units (Cluster 3) is conducted from each of these secondary units, and so on, until we reach the final stage of breakdown for the sample units, when we sample every member of those units.

9.3.2 Non-probability Sampling

Some of the non-probability sampling techniques are commonly used explicitly in cases where it is not possible to use the probability sampling. The major difference is that in non-probability techniques, the extent of bias in selecting the sample is not known. This makes difficult to say anything about the representativeness or accuracy of the sample. Nevertheless, if done carefully, some of these are good approximation of probability sampling.

Convenient sampling

It refers to the collection of information from members of the population who are conveniently available to provide it. It involves picking up any available set of respondents convenient for the researcher to use. For example, suppose

■ Convenience sampling refers to the collection of information from members of the population who are conveniently available to provide it.

a manager of the restaurant situated in Shah Alam wants to know what type of drinks generally people prefer. In this case the most convenient way to find the sampling unit is to visit the different restaurants in Malaysia. Suppose a researcher is interested to measure the service quality of CIMB bank. The most convenient way to find the sample (i.e., the customers of CIMB bank) is to visit branches of CIMB bank. Similarly, suppose a student of MBA is conducting a management research to measure the shopping values among the Malaysian youth. For his/her convenience, he/she may choose majority of the sample within the university. These are some of the examples of convenience sampling. The convenience sampling is very often used by researchers in order to cover the large number of survey quickly and cost effectively. However, it suffers from selection bias because the individuals surveyed are often different from the target population and thus may not be a true representative of the population for the study in hand.

Judgement sampling

■ Judgement sampling involves the choice of subjects who are most advantageously placed to provide the information required.

A judgement sample, sometimes referred to as a purposive sample, involves selecting elements in the sample for a specific purpose. It involves the choice of subjects who are most advantageously placed or in the best position to provide the information required. Judgement sample might be a group of experts with knowledge about a particular problem or issue. For example, physicians who specialize in treating diabetes might be interviewed in a survey to gain insight about the most effective ways to convince diabetics to adopt good diets and proper exercise. Similarly, suppose you wanted to interview incentive travel organizers within a specific industry to determine their needs or destination preferences. You may find that not only they are few in numbers but also extremely busy and thus, may well be reluctant to participate in the interview. Relying on the judgement of some knowledgeable experts may be far more productive in identifying potential interviewees than trying to develop a list of the population in order to randomly select a small number.

Quota sampling

■ In quota sampling, groups are adequately represented in the study through the assignment of a quota based on known characteristics.

Quota sampling is a non-probability sampling technique, wherein certain groups are adequately represented in the study through the assignment of a quota based on known characteristics. Quota sampling could be either proportional or disproportional.

In proportional quota sampling we represent the major characteristics of the population by sampling a proportional amount of each. Suppose the total sample size of 200 is to be targeted based on the strata "Gender". Let us say that the population has approximately 55% women and 45% men. In this case we will continue sampling until we get those percentages. So, if we have already got the 90 men for our sample, but not the 110 women, we will continue to sample women; but even if legitimate men respondents come along, we will not sample them because we have already "met the quota for men". The problem here (as in much purposive sampling) is that we have to decide the specific characteristics on which we will base the quota. Disproportional quota sampling is a bit less restrictive. In this method, we

specify the minimum number of sampled units we want in each category. Here, we are not concerned with having numbers that match the proportions in the population. Instead, we simply want to have enough to assure that we will be able to talk about even small groups in the population. This method is typically used to assure that smaller groups are adequately represented in the sample.

Snowball sampling

The snowball sampling is also known as the sampling by reference where one respondent is being used to generate names of others.. In this case, the researcher uses the initial respondents (who are chosen using probability method) to help identify the other respondents in the target population. This technique is used when the population being sought is a small one, and chances of finding the elements of the population by the traditional method are difficult. The university being the non-smoking zone, the management of university is interested to find out the smoking habits among the students. In this case, the target population is all the students of the university irrespective of their gender, who are in the habit of smoking. But it is not so easy on the part of the researcher to find out the elements of the population by common method. He can get the list of students currently studying in the university from the administrative office, but not the list of students who are smokers. One of the best ways to get the sample in this case is through reference by initial respondents whom the researcher probably knows as the smokers.

■ The snowball sampling is also known as the sampling by reference.

9.4 Determining the Sample Size

Determining sample size is a very important issue because samples that are too large may waste time, resources and money, while samples that are too small may lead to inaccurate results. To determine the sample size three pieces of information are required. They are: (i) the degree of confidence necessary to estimate true value, (ii) the precision of the estimate and (iii) the amount of true variability present in the data.

9.4.1 Level of Precision

Precision refers to how close our estimate is to the true population characteristics. We estimate the population parameter to fall within a range based on the sample estimate. For example, let us say that from a study of simple random sample of 40 of the total 200 workers who are involved in assembling the printed circuit board (PCB) in a PCB Assembly Company, we find that the daily average number of assembled PCB per worker is 50 ($\bar{X}=50$). Let us say that the true average number of assembled PCB for the entire population would lie somewhere between 40 to 60 PCBs. In saying this, we offer an interval estimate within which we expect the true population mean production to be ($= 50 \pm 10$). The narrower this interval, greater is the precision. For instance, if we are in position to estimate the daily assembling rate for the population somewhere between 45 to 55 ($= 50 \pm 5$), then we say that we have comparatively more precision. That is, we would now estimate

■ Precision implies how close our estimate is to the population parameter.

the population mean to lie within narrower range, which in turn means that we would now estimate with greater precision.

9.4.2 Variability in Data

Precision is a function of the range of variability in the sampling distribution of the sample mean. If we take a number of different samples from a population and take the mean of each of these, we shall usually find that they are all different, normally distributed and have a dispersion associated with them. The smaller the dispersion or variability, the greater the probabilities that the sample mean will be closer to the population mean. However, there is no need to take several different samples to estimate the variability. Just by taking one sample of say 30 subjects from the population, we can estimate the variability of the sampling distribution of the sample mean. This variability is known as standard error ($S_{\bar{x}}$) and is calculated as:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

Where S is the standard deviation of the sample, n is the sample size and $S_{\bar{x}}$ is the standard error or the extent of precision offered by the sample. Two points to be noted here:

- (i) In order to reduce the standard error ($S_{\bar{x}}$), one needs to increase the sample size at a given standard deviation of the sample (S).
- (ii) Smaller the variation in population, the smaller the standard error ($S_{\bar{x}}$), which in turn implies that sample size need not be large.

Thus, we need the greater precision if we want our sample results to closely reflect the characteristics of the populations. The greater the precision required, the larger is the sample size needed, particularly when the variability in the population is large.

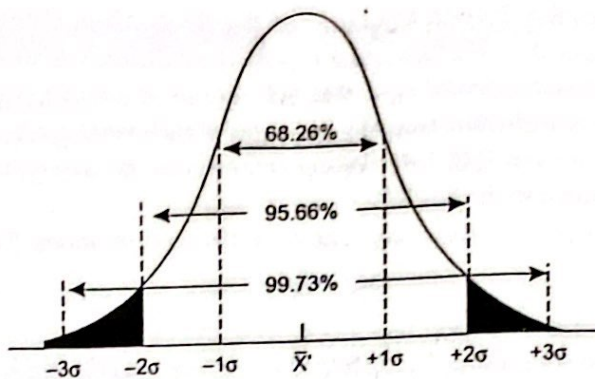
9.4.3 Level of Confidence

The *confidence* or *risk level* is based on ideas encompassed under the Central Limit Theorem. The key idea encompassed in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value, with some samples having a higher value and some obtaining a lower score than the true population value. In a normal distribution, approximately 95% of the sample values are within two standard deviations (2σ) of the true population mean (μ).

In other words, this means that, if a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision specified earlier (see Figure 9.1). There is always a chance that the sample you obtain does not represent the true population value. Such samples with extreme values are represented by the shaded areas in Figure 9.1. This risk is reduced for 99% confidence levels and increased for 90% (or lower) confidence levels.

■ The smaller the dispersion or variability, the greater the probabilities that the sample mean will be closer to the population mean.

■ The level of confidence denotes how certain we are that our estimates will really hold true for the population.

Figure 9.1
Normal curve

9.4.4 Sample Data, Precision and level of Confidence in Estimation

Precision and confidence play vital role in sampling as we use sample data to draw inferences about the entire population. Because the point estimate provides no measure of possible error, to be fair we do an interval estimation to ensure a relatively accurate estimation of the population parameters.

Let us say, we want to estimate the daily mean (average) assembled PCB for the entire population (workers) of 200. From a sample of 40 workers chosen randomly, let us say the sample mean $\bar{X} = 50$ and the sample standard deviation $S = 8$. $\bar{X}$, the sample mean is the point estimate of μ , the population mean. We shall construct the confidence interval around $\bar{X}$ to estimate the range within which μ would fall.

The standard error $S_{\bar{X}}$ and the level of confidence determine the width of the interval and calculated by using the following formula.

$$\mu = \bar{X} \pm K S_{\bar{X}}$$

Where, K is the t statistics for the level of confidence desired.

$$\text{For our example, } S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{8}{\sqrt{40}} = \frac{8}{6.324} = 1.265$$

From the table of critical values for t as given in the appendix at the end of the book, we know that:

- For 90% level of confidence, the K value is 1.645.
- For 95% level of confidence, the K value is 1.96.
- For 99% level of confidence, the K value is 2.576.

Thus, at 90% level of confidence, $\mu = 50 \pm 1.645(1.265)$, i.e., $\mu = 50 \pm 2.081$ - indicating that μ would fall between 47.919 and 52.081. It indicates that with a sample size of 40, we have 90% confidence that the true population mean (average number of assembled PCBs) for all the workers would fall between 47.919 and 52.081. Similarly if want to be relatively more confident say 99% without increasing the sample size, we need to sacrifice the precision. In this case, $\mu = 50 \pm 2.576(1.265)$, i.e., $\mu = 50 \pm 3.259$ - indicating that the true value of μ would lie between 46.741 and 53.259. The width of the estimation has increased and thus we are relatively less precise in estimating the population parameter. Thus there is a trade-off between *precision* and *level of confidence*.

There is a trade-off between precision and level of confidence.

9.4.5 Sample Size

Let us suppose, the Branch Manager of the Shah Alam CIMB bank wants to be 95% confident that the expected monthly withdrawals in the bank will be within a confidence interval of $\pm RM 400$. Let us say that a study of sample of banks customers indicates that the average withdrawal made by them have a standard deviation of RM 2800. What sample size is needed under the above criteria and information given?

As we know, population mean can be estimated by using the formula:

$$\mu = \bar{X} \pm K S_{\bar{X}}$$

Since the confidence level needed is 95%, the corresponding K value is 1.96 (t table). The interval estimate of RM 400 will have to encompass a dispersion of (1.96 x standard error). That is,

$$\begin{aligned} 400 &= 1.96 \times S_{\bar{X}} \\ \Rightarrow S_{\bar{X}} &= 400/1.96 = 204.08 \end{aligned}$$

We know that

$$\begin{aligned} S_{\bar{X}} &= \frac{S}{\sqrt{n}} \\ \Rightarrow 204.08 &= \frac{2800}{\sqrt{n}} \\ \Rightarrow \sqrt{n} &= \frac{2800}{204.08} = 13.72 \\ \Rightarrow n &= (13.72)^2 = 188.23 \approx 188 \end{aligned}$$

The sample size needed is 188. Let us further assume that the Shah Alam branch of CIMB bank has total customers of only 185. This means we cannot sample 188 customers. In this case we need to apply the correction formula in order to estimate the sample size under the same conditions of precision ($\pm RM 400$) and confidence level (95%).

The correction formula is given as:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Where N is the size of the population, n is the sample size to be estimated, $S_{\bar{X}}$ is the standard error of estimate of mean and S is the standard deviation of the sample mean.

Applying the correction formula, we get

$$\begin{aligned} 204.08 &= \frac{2800}{\sqrt{n}} \times \sqrt{\frac{185-n}{184}} \\ \Rightarrow \frac{\sqrt{n}}{\sqrt{185-n}} &= \frac{2800}{204.08 \times \sqrt{184}} = 1.01146 \\ \Rightarrow \frac{n}{\sqrt{185-n}} &= (1.01146)^2 = 1.023051 \\ \Rightarrow n &= 187.1201 - 1.023051(n) \\ \Rightarrow 2.023051(n) &= 187.1201 \\ \Rightarrow n &= 92.492 \approx 92 \end{aligned}$$

Thus, we need approximately 92 customers out of total 185 customers.

9.5 Factors Affecting the Decision on Sample Size

It is not the formula alone that determines the sample size in actual business research. Sampling in practice is based on science, but it is also an art. The basic assumptions made while computing sample size through the use of formula are sometimes not met in practice. Roscoe (1975) proposes the following rules of thumb for determining the sample size:

- Sample size larger than 30 and less than 500 are appropriate for most research.
- Where samples are to be broken into sub-samples such as male/female, Malay/Chinese/Indian, etc., a minimum sample size of 30 for each category is necessary.
- In multivariate data analysis (including multiple regression analysis), the sample size should be several times (preferably 10 times or more) as large as the number of variables in the study.
- For simple experimental research with tight experimental controls (matched pairs, etc.), successful research is possible with samples as small as 10 to 20 in size.

9.6 Types of Error in Business Research

The research in social science or management science is not free from error. Error in research is inevitable whether it is done by a professor from Aston Business School or by an average student of any university in Malaysia. However, the extent of error cannot be the same. For example, in the above case the error can be minimal if it is done by an experienced professor from Aston University and error might be quite large in the latter case. There are two types of errors that are found in any social or management science research.

9.6.1 Sampling Error

This is the error which occurs due to the selection of some units and non-selection of other units into the sample. It is controllable if the selection of the sample is done in random, unbiased way. In other words, if the probability sampling is used, it is possible to control this error. In general, *sampling error reduces as the sample size increases.*

■ Sampling error reduces as the sample size increases.

9.6.2 Non-sampling Error

This is the effect of various errors in doing the study, by the interviewers, data entry operator or the researcher himself. Handling a large quantity of data is not an easy job, and errors may creep in at any stage of research. The data entry person may interchange the column of yes and no responses while entering and compiling the data, or the interviewer may cheat by not filling up the questionnaire in the field and instead, fudge the data. Or the respondent may say one thing, but another may be recorded by mistake. These errors are usually the function of sample size. That is, *the larger the sample size, the larger the non-sampling error.*

■ The larger the sample size, the larger the non-sampling error.

The total error in research is the sum of above two errors.

$$\text{Total Error} = \text{Sampling Error} + \text{Non-sampling Error}$$

The sampling error can be estimated in the case of probability sampling, but not in the case of non-probability sampling. Non-probability sampling can be controlled through hiring better field workers, qualified data entry persons, and good control procedures at every stage of research.

Chapter Summary

One of the important aspects of research design is the sampling design. A researcher has to take decision on both the sampling technique to be used and the sample size that is needed. Both are important to establish the representativeness of the sample for generalizability.

The sampling technique can be broadly divided into probability sampling and non-probability sampling. In probability sampling technique, a sample is being selected using random selection so that each element of the population has a known chance of being selected. In non-probability sampling techniques, the extent of bias in selecting the sample is not known. This makes difficult to say anything about the representativeness or accuracy of the sample. Though non-probability sampling designs have limitations in terms of generalizability, they are often the only designs available for certain types of investigations.

To determine the sample size three pieces of information are required: (i) the degree of confidence necessary to estimate the true value, (ii) the precision of the estimate and (iii) the amount of true variability present in the data. Precision refers to how close our estimate is to the true population

characteristics. The smaller the dispersion or variability in data, the greater the probabilities that the sample mean will be closer to the population mean. The level of confidence denotes how certain we are that our estimates will really hold true for the entire population. There is a trade-off between precision and level of confidence. If we want more precision, or more confidence, or both, the sample size needs to be increased unless of course, there is very little variability in the population itself. However, if the sample size cannot be increased say because of budget constraint, the only way to maintain the same level of precision would be by forsaking the confidence with which we can predict the population.

The total error in research is composed of sampling error and non-sampling error. The sampling error is due to the selection of some units and non-selection of other units into the sample. This can be controlled by choosing the appropriate probability sampling design. On the other hand, the non-sampling error occurs because of the occurrence of errors in any stage of the research process, such as errors by fieldworkers, data entry person or the researcher himself/herself.

Discussion Questions

- 1 Non-probability sampling might be preferred to probability sampling designs in some cases. Explain with examples.
- 2 There is a trade-off between precision and level of confidence. Explain.
- 3 With examples, explain the difference between single stage and multistage cluster sampling. What are the advantages and disadvantages of cluster sampling?
- 4 There is a trade-off between sampling and non-sampling error. Discuss.
- 5 Give an example of research study where the stratified random sampling can be applied. Also explain the procedure of drawing proportionate and disproportionate random sampling with hypothetical data.
- 6 Write short notes on the following sampling designs:
 - (a) Convenience sampling
 - (b) Snowball sampling
 - (c) Quota sampling
- 7 A fast-food company in Malaysia wants to determine the average number of times that fast-food customers visit fast-food restaurants per week. They have decided that their estimate needs to be accurate within plus or minus one-tenth of a visit, and they want to be 95% sure that their estimate does differ from true number of visits by more than one-tenth of a visit. Previous research has shown that the standard deviation is 0.7 visits.
 - (a) What is the required sample size?
 - (b) How many more sample will be required if the company wants to be 99% confident in estimation with the same level of precision and standard deviation?
 - (c) What is the sample size requirement if the company wants the estimation within plus or minus one-eighth of a visit with 95% level of confidence?